The Correlation Between Education Attainment, Employment, and Income

Zeeshan Pervaiz, Sarvani Dantuluri, Cheyenne Peterson, Rob Dewan

DEV10 Data Professional Project Capstone

February 3rd, 2023

Table of Contents

Introduction

The objective of this comprehensive report is to analyze and compare various datasets relating to Education, Income, and Unemployment Rates in the United States. Our aim is to gain a deeper understanding of the relationship between these factors and to answer questions that arose during our research. We carefully examined the data over the years to observe any trends or changes that have occurred. Our research focused on the following questions: What is the average per capita income for individuals with different levels of education? How does the average household income in each state compare to the national average? Is there a correlation between household income and educational attainment? What is the unemployment rate among individuals with varying levels of education? How has educational attainment evolved over time in the United States and how has the income changed accordingly? Through our extensive analysis of the available data, we aim to provide insights and a thorough understanding of the complex relationships between Education, Income, and Unemployment Rates in the United States. Our hypothesis is that there is a relationship between Education, Income, and Unemployment Rates in the United States, and that this relationship has evolved over time.

Data

In this project, there were a total of four datasets that were gathered to study the relationship between education attainment, employment, individual income, and household income. These datasets were primarily obtained from the Bureau of Labor and Statistics and the U.S. Census Bureau. The data used was converted into a csv file and later transformed into a dataframe for the purpose of creating a SQL database and visualizations for the dashboard. The following is a list of the four datasets with details, and the sources are listed in the references.

Name	**Description**	**Columns/DataTypes**	**Shape**
Cpsaat_2015_2021 SQL Table: Employment	The employment status of the civilian noninstitutional population 25 years and over by educational	1. Civilian noninstitutional population**(int)** 2. Civilian labor force**(int)** 3. Participation rate**(float)** 4. Employed**(float)** 5. Employment-population ratio**(float)** 6. Unemployed**(float)** 7. Unemployment rate**(float)** 8. Sex**(str)** 9. Race**(str)**	(392,10)

	attainment, sex, and race.	10. Year**(int)**	
Table-A2 SQL Table: Income	Households by total money income and race of householder: 1967 to 2021	1. Race**(str)** 2. Year**(int)** 3. Number(Thousands)**(float)** 4. Total Percent Distribution**(float)** a. Under $15000**(float)** b. $15000 to $24000**(float)** c. $25000 to $34999**(float)** d. $35000 to $49000**(float)** e. $50000 to $74999**(float)** f. $75000 to $99999 **(float)** g. $100000 to $149999**(float)** h. $150000 to $199999**(float)** i. $200000 and over**(float)** 5. Median Income Estimate**(float)** 6. Median Income Margin of Error**(float)** 7. Mean Income Estimate**(float)** 8. Mean Income Margin of Error**(float)**	(437,17)
Tabn102-30 SQL Table: Household	Median household income by state: selected years, 1990 through 2019	1. State **(str)** 2. Median Household Income by State: 1990**(int)** 3. Median Household Income by State: 2000**(int)** 4. Median Household Income by State: 2005**(int)** 5. Median Household Income by State: 2010**(int)** 6. Median Household Income by State: 2014**(int)** 7. Median Household Income by State: 2015**(int)** 8. Median Household Income by State: 2016**(int)** 9. Median Household Income by State: 2017**(int)** 10. Median Household Income by State: 2018**(int)** 11. Median Household Income by State: 2019**(int)**	(52,11)
Taba-3 SQL Table: Degrees	Mean earnings of workers 18 years and over, by educational attainment, race, and sex: 1975-2020	1. Race**(str)** 2. Sex**(str)** 3. Year**(int)** 4. Total Mean**(float)** 5. Total Number with Earnings**(float)** 6. Not a High School Graduate Mean**(float)** 7. Not a High School Graduate Numbers with Earnings**(float)** 8. High School Graduate Mean**(float)** 9. High School Graduate Numbers with Earnings**(float)** 10. Some College/Associate's Degree Mean**(float)**	(872,15)

		11. Some College/Associate's Degree with Earnings**(float)** 12. Bachelor's Degree Mean**(float)** 13. Bachelor's Degree Numbers with Earnings**(float)** 14. Advanced Degree Mean**(float)** 15. Advanced Degree Numbers with Earnings**(float)**	

Data Processing

The data cleaning process involved downloading the data into Excel and converting it to a CSV file, which was then transferred to a group SQL database using Databricks. There were 5 SQL tables, including degrees, employment, household, income, and household_abbr. A config.py file was created to import the tables from SQL to vs code.

For the degrees dataset, certain columns were dropped and the remaining ones were renamed. The mean and number of earnings columns were converted to a float data type, the year column was converted to an int data type, and the race and sex columns were converted to an object data type.

The degrees_income dataset was created by selecting specific columns from the degrees dataset to create a visualization of per capita income average by education level and gender. The sex column was filtered to show only male and female, and the race column was filtered to "total."

The degrees_income_race dataset was created in a similar manner to the degrees_income dataset, with the difference being that the sex column was filtered to show "both sexes" and the race column was filtered to "total."

The degrees_income_year dataset was created by selecting specific columns from the degrees dataset to create a visualization of per capita income by education level from 2010 to 2019. The year column was filtered to show only from 2010 to 2019, and the sex and race columns were filtered to show "both sexes" and "total."

For the Employment_status dataset, the columns for participation rate, employment-population ratio, unemployment rate, employed, and unemployed were converted to a float data type, and the year, civilian noninstitutional population, and civilian labor force columns were converted to an int data type.

The degree_employment_status dataset was created by selecting specific columns from the employment_status dataset to create a comparison of unemployment rates by education level. The education attainment was filtered to exclude "Some college or associate degreeTotal" and "Bachelor's degree and higherTotal."

The degree_employment_rates dataset was created in a similar manner to the degree_employment_status dataset, but only filtered the year column to 2020.

For the household_abbr and household datasets, the header index was dropped and the column names were renamed. The Median Household Income by State columns were converted to a float data type.

For the income dataset, the index header was removed and replaced with column names. The columns were renamed as Race, Year, Number(Thousands), Total Percent Distribution, Under $15000, $15000 to $24999, $25000 to $34999, $35000 to $49999, $50000 to $74999, $75000 to $99999, $100000 to $149999, $150000 to $199999, $200000 and over, Median Income Estimate, Median Income Margin of Error, Mean Income Estimate, and Mean Income Margin of Error. The year column was changed to an integer type and the rest of the columns were converted to a float type.

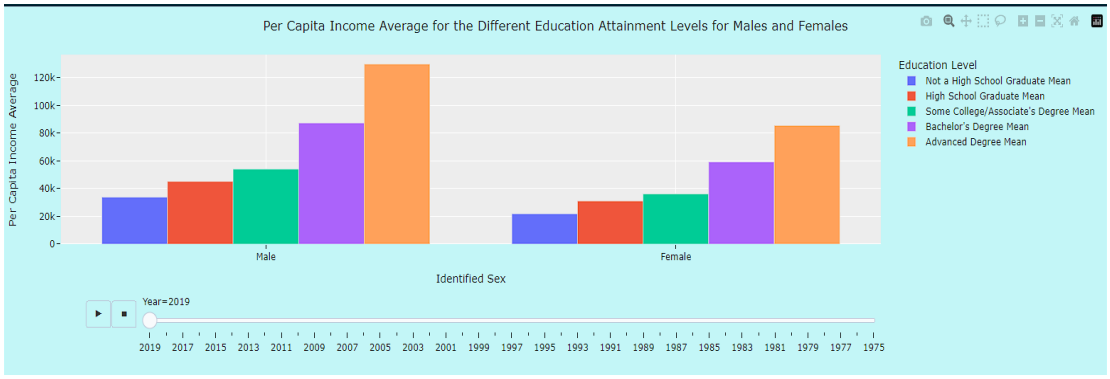
To create the "Estimated Average Income for 2015-2019 in the US" visualization, a new data frame was created from the income dataset, selecting the year, mean income estimate, median income estimate, and race columns. The data was filtered to show years from 2015 to 2019 and race to "all races". The mean income estimate column was converted to a float type and the race column was converted to a string type.

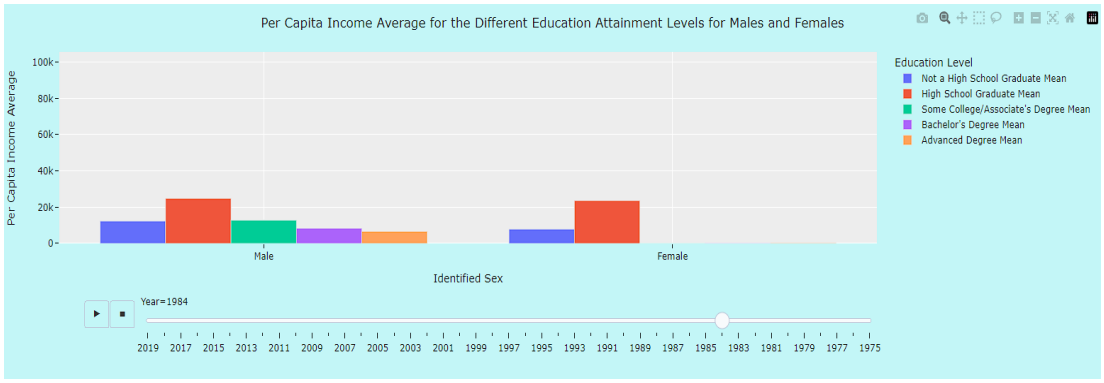
For the "per capita income for the identified race" visualization, a new data frame was created using the year, mean income estimate, median income estimate, and race columns from the income dataset. The race column was filtered to exclude "total". The mean and median income estimate columns were converted to float types and the race column was converted to a string type.

For the household_degree2 dataset, the household dataset was transposed to correct the indexing, then the index was renamed and reset. A new column for the years 1990, 2000, 2005, 2010, 2014-2019 was created. The household and degrees datasets were merged on the year column to create a visualization comparing household income across races for selected years. The household_degree2 dataset was created from the merged dataset, filtered to show years from 2015 to 2019 and race not equal to "total", and with the sex column set to "both sexes". The data types for the degree means were all changed to float.

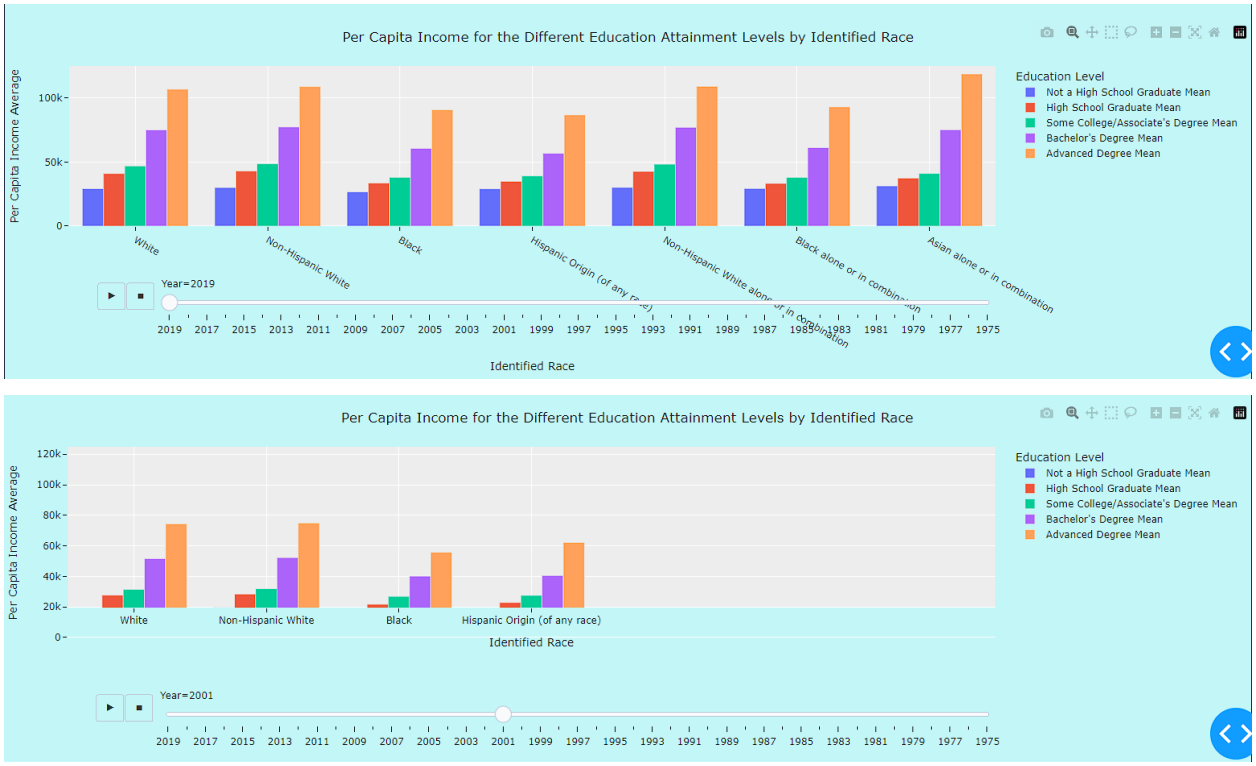
Data Analysis

The first question of our analysis focused on exploring the relationship between per capita income and educational attainment. Our findings indicated that, as expected, individuals with advanced degrees tend to earn higher incomes compared to those with lower levels of education. However, if you move the year slider, it is intriguing to note that this trend was not observed 40 years ago, when individuals with only high school degrees were earning more than those with higher education degrees. This could be attributed to the fact that women were not even surveyed at that time, suggesting that they were not actively participating in the workforce or pursuing higher education.





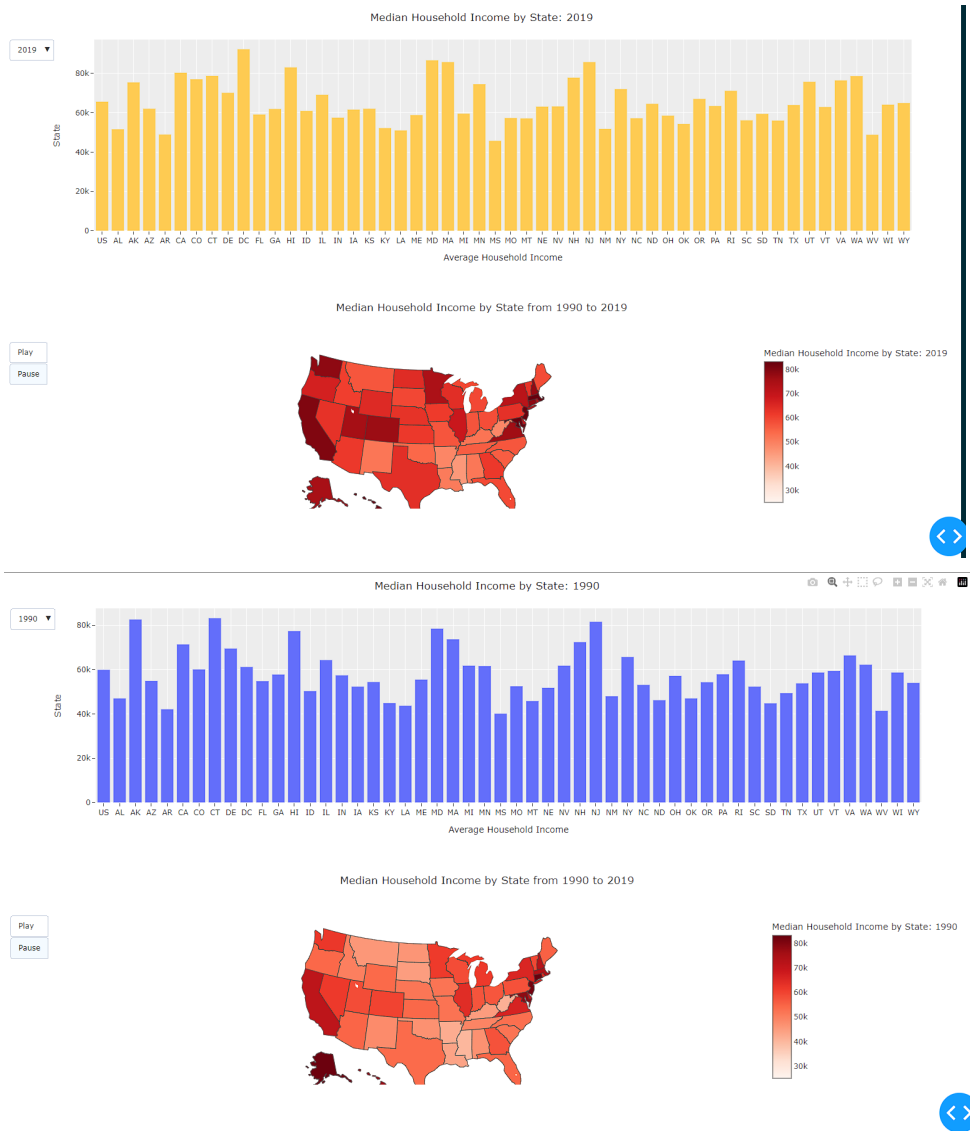
Additionally, our analysis revealed a persistent gender gap in income, with men earning higher incomes regardless of their educational attainment. An examination of racial demographics showed that individuals of Asian descent tend to earn the highest incomes among the racial groups. It is noteworthy that this demographic was not recorded by the census bureau until 2002, with regards to educational earnings.



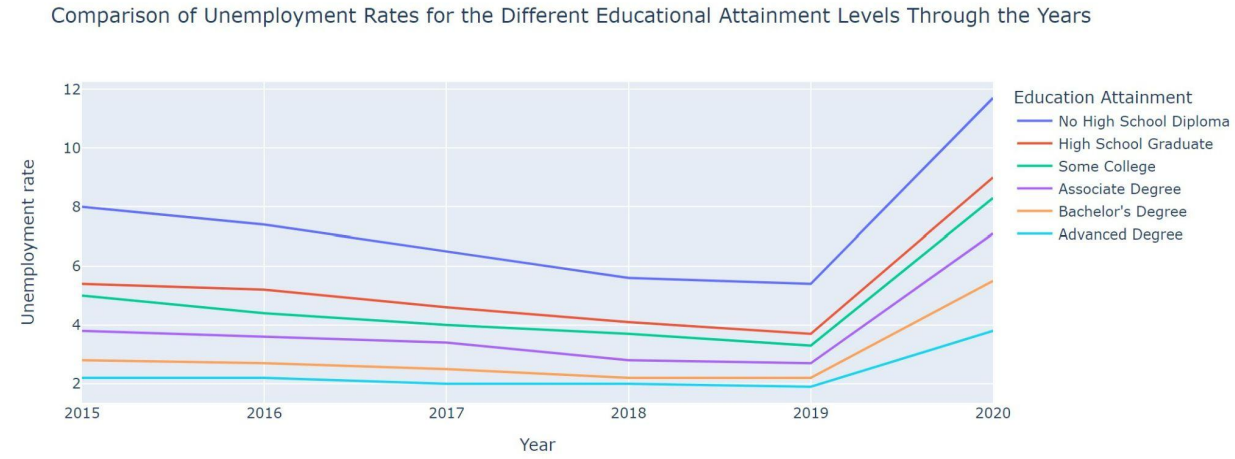
The second question of our analysis aimed at comparing average household income across different states in the US from 1990 to 2019. To uncover the answers, I created both a bar graph that compared state average household incomes to the US average as a whole and a choropleth map that highlights the differences by color variances throughout the years. Our findings showed that there are several states whose average household income is significantly higher than the national average. In particular, Washington DC was observed to have higher household incomes in 2019 with Maryland historically having the highest average income in earlier years. Some states like Arizona, Mississippi, and West Virginia had average incomes that were well below the national average. Which for those three states, has not changed significantly since 1990. Lastly, there are states that were very high up in household income averages in the early 90s but somehow declined immensely in the coming decades up until 2019 like Alaska and Connecticut. There are many factors that could contribute to the differences in average household income across different states in the US. Some of these factors include:

1. Cost of living: The cost of living in different states can vary significantly, which can impact household incomes.
2. Economic conditions: The strength of the local economy, unemployment rate, and availability of jobs can all impact household incomes.

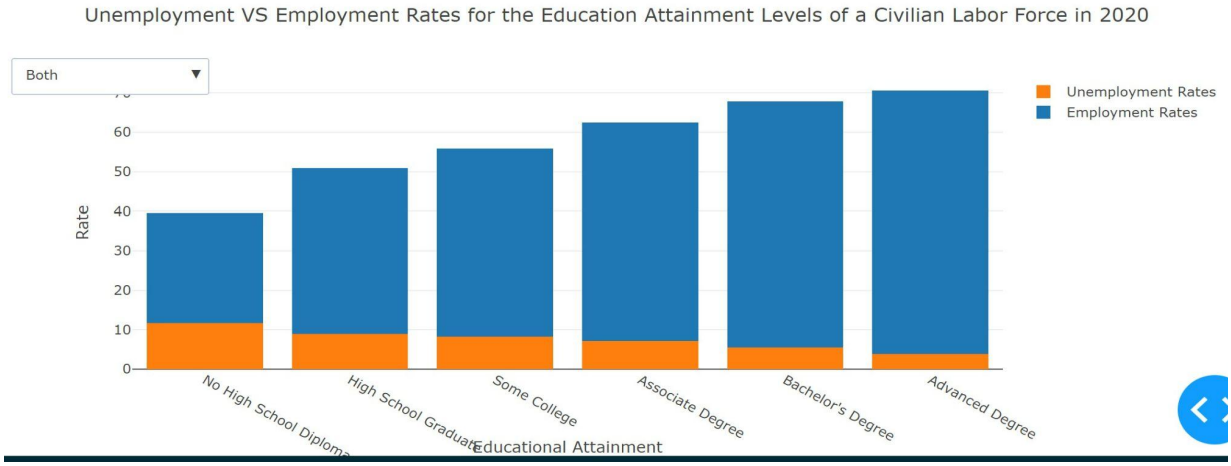
3. Education level: States with higher levels of education tend to have higher average household incomes.
4. Industry composition: The types of industries that are prevalent in a state can also impact household incomes. For example, states with a strong technology sector may have higher household incomes.
5. Taxation policies: Different states have different tax policies, which can impact household incomes by either increasing or decreasing the amount of disposable income people have.
6. Migration patterns: Changes in population due to migration can also impact average household incomes in a state.



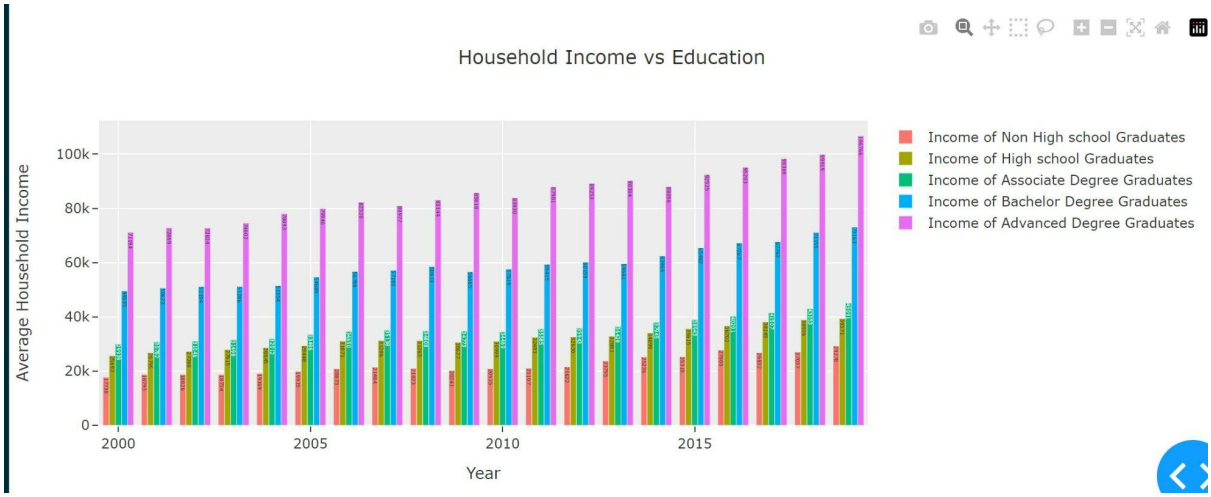
For Exploratory Question #3, two visualizations were created to investigate the unemployment rate for various levels of education. The first visualization shows the unemployment rate for each educational attainment from 2015 to 2020. It is evident that those without a high school diploma have the highest unemployment rate compared to those with other levels of education. Meanwhile, those with advanced degrees have the lowest unemployment rate. A notable observation is that the unemployment rate for each educational attainment level rose significantly from 2019 to 2020, which could be attributed to inflation and the early stages of the COVID-19 pandemic.



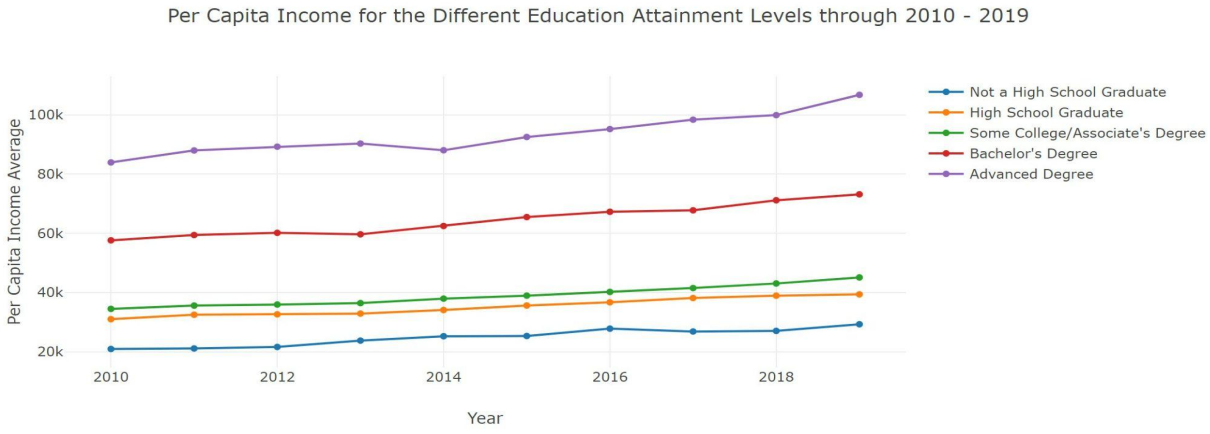
The second graph depicts the comparison of unemployment and employment rates for civilian workers in 2020, classified by their educational attainment level. It is evident that those with an advanced degree have a higher rate of employment compared to those without a high school diploma. It also seems that those without a high school diploma have a higher rate of unemployment compared to those with an advanced degree. In general, the trend indicates that individuals with more experience or higher education are more likely to be employed than those with less experience.



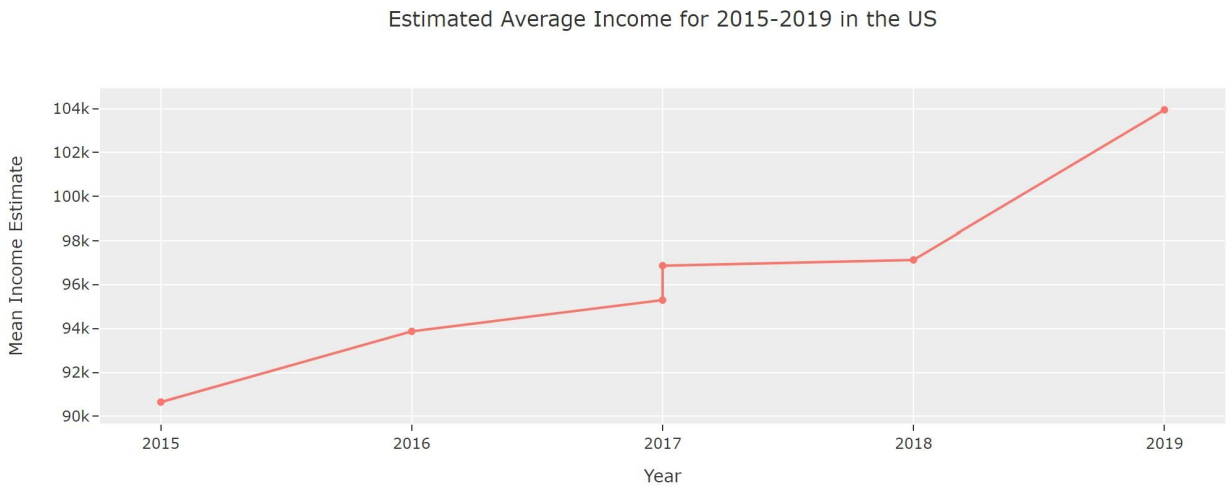
The visualization created for explanatory question #4 shows the average household income for different educational degrees from 2000 to 2019. It is clear that those with advanced degrees have higher household income compared to those with lower degrees, and they also experience the largest increase in income, rising from $71k to $106k. A steady income difference of $11k is observed for non-high school graduates. However, it is noteworthy that high school graduates have a larger income gap of $24k compared to those with Associate Degrees, who have a difference of $15k.



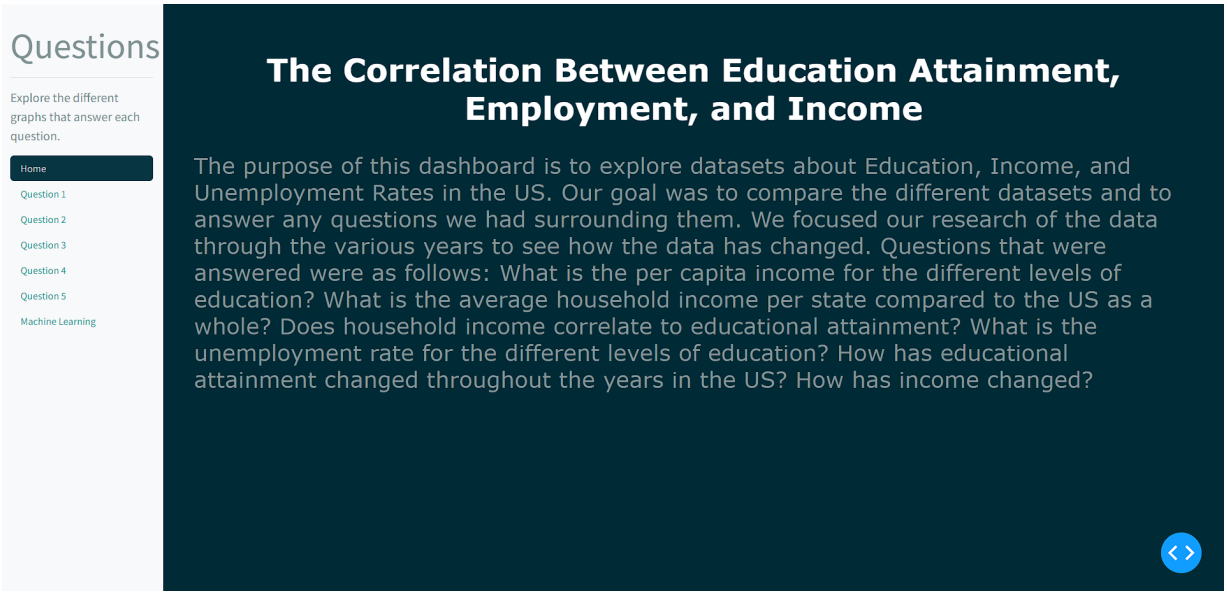
Two visualizations were created for the fifth explanatory question to demonstrate the evolution of educational attainment and income in the US from 2010 to 2019. The first visualization depicts the relationship between individual income and level of education. It is clear that those with advanced degrees tend to earn more than those with lower levels of education, with non-high school graduates earning the least. Interestingly, the income line for individuals with advanced degrees fluctuates the most compared to other lines.



For the second visual, the trend of average income in the US from 2015 to 2019 was plotted. It showed a significant increase from 2015 to 2019, starting at roughly $91,000 in 2015 and reaching $104,000 in 2019. There was a noticeable deviation in the trend in 2017, where the data was recorded twice, leading to a stable income from 2017 to 2018. The overall trend indicated that the mean income was rising as the years went by.



Dashboard



To present our visualizations, we decided to utilize Dash dashboard via Visual Studio Code. Before writing any code on our python file that would be used for the dashboard, we created nearly all of our visualizations on a Jupyter Notebook to explore what we did and did not want when it came to answering the questions. We created a multipage dashboard that filters our visualizations based on our exploratory questions with a separate tab for machine learning. The code starts by importing the required libraries such as dash, dash_bootstrap_components (dbc), and plotly. The dash.Dash() function creates a new instance of the Dash application and sets the name of the instance to name and sets the external stylesheet to the solar theme from dash_bootstrap_components.

```
# Dashboard
app = dash.Dash(__name__, external_stylesheets=[dbc.themes.SOLAR])

SIDEBAR_STYLE = {
    "position": "fixed",
    "top": 0,
    "left": 0,
    "bottom": 0,
    "width": "16rem",
    "padding": "2rem 1rem",
    "background-color": "#f8f9fa",
}

# the styles for the main content position it to the right of the sidebar and
# add some padding.
CONTENT_STYLE = {
    "margin-left": "18rem",
    "margin-right": "2rem",
    "padding": "2rem 1rem",
}

sidebar = html.Div(
    [
        html.H2("Questions", className="display-4"),
        html.Hr(),
        html.P(
            "Explore the different graphs that answer each question.", className="lead"
        ),
        dbc.Nav(
            [
                dbc.NavLink("Home", href="/", active="exact"),
                dbc.NavLink("Question 1", href="/page-1", active="exact"),
                dbc.NavLink("Question 2", href="/page-2", active="exact"),
                dbc.NavLink("Question 3", href="/page-3", active="exact"),
                dbc.NavLink("Question 4", href="/page-4", active="exact"),
                dbc.NavLink("Question 5", href="/page-5", active="exact"),
                dbc.NavLink("Machine Learning", href="/page-6", active="exact")
            ],
            vertical=True,
            pills=True,
        ),
    ],
    style=SIDEBAR_STYLE,
)
```

Next, two dictionaries named SIDEBAR_STYLE and CONTENT_STYLE are defined to store the CSS styles for the sidebar and main content of the dashboard respectively. Afterwards, a new div element is created and stored in the "sidebar" variable. This div element contains the elements that will be displayed in the sidebar of the dashboard. The div element contains a header "Questions", a horizontal rule (hr), a paragraph explaining the purpose of the dashboard, and a navigation bar. The navigation bar contains links to different pages in the dashboard. The style of this div is set to the SIDEBAR_STYLE. The content variable is also defined as a div element with the id "page-content". This div element will store the main content of each page in the dashboard and its style is set to the CONTENT_STYLE.

I defined the layout of the dashboard as a div element containing three components: the location component with the id "url", the sidebar div, and the content div. The location component is used to determine the current URL path and is used to update the content displayed in the content div. Finally, a callback function is defined that takes the pathname of the current URL as an input and returns the children of the content div. The function uses an if-elif statement to check the pathname and returns different content based on the pathname. If the pathname is "/", it returns a header and

a paragraph explaining the purpose of the dashboard. If the pathname is "/page-1", it returns a header and two graphs showing the per capita income for different levels of education. Similarly, other pages return different content based on their corresponding pathnames. After creating the bones of the dashboard, I inserted our combined created visuals/graphs and created titles for each tab on the sidebar.

Machine Learning

For our Machine Learning section we were aiming to determine if we could predict the unemployment rate based on a number of other features in the dataset. We utilized the Unemployed, Employed, and Employment-population ratio features to predict the unemployment rate. We were a bit curious to see which regression model would have the best score on the dataset so we tested 4 of them. We tested Linear Regression, Decision Tree Regression, K Nearest Neighbors (KNN) and Ridge Regression. We tuned hyperparameters for the KNN model and the Ridge Model. For KNN we set the neighbors value to 4. For Ridge we defined the hyperparameters as alpha = 0.5, normalize = False, tol = 0.001, solver ='auto', random_state = 42. The scores obtained from our models are summarized below.

Model	Score
Linear Regression	0.5414024159979208
Decision Tree Regression	0.7767003702629047
K Nearest Neighbors	0.5733274263535004
Ridge Regression	0.5418758699091106

Generally speaking, Regression model scores are used to evaluate the performance of the model, which is a statistical technique used to predict a target variable (unemployment rate) based on one or more input features (Unemployed, Employed, and Employment-population ratio). The scores help to quantify the accuracy of the model's predictions. Based on the scores obtained, the Decision Tree Regression model had the highest score whereas the other 3 had similar lower scores. We can say that the DT model was the most accurate in predicting unemployment rate based on the input features.

Results

The average income of individuals has seen an upward trend over the years. This can be attributed to a number of factors, including increased access to higher education and a growing economy. Education has a significant impact on one's income, with individuals who possess a degree typically earning more than those who do not. This disparity is even more pronounced for those with advanced degrees. It's worth noting that the average state income has undergone significant changes from 1990 to 2019, reflecting broader economic trends and shifts in the job market. The working population has continued to grow, but so have unemployment rates. Despite this, those with advanced degrees are still more likely to be employed than those with only a GED. Furthermore, the average income for people with advanced degrees has increased greatly over time, while income for individuals with no GED has remained relatively stagnant.

This information can be useful for a business in a number of ways. Firstly, by understanding that the average income has increased over the years, a business can adjust its pricing and compensation strategies to match the current market trends. Additionally, businesses can focus on hiring individuals with higher levels of education, as these individuals typically have higher incomes and are more likely to be employed. By doing so, businesses may be able to attract and retain high-quality employees, which can improve productivity and overall performance. Furthermore, by understanding the changes in average state income and the total working population, businesses can identify potential markets for growth and expansion. For example, if the working population is increasing and average income is rising in a particular state, that could be a good indicator for businesses to expand their operations into that area. Ultimately, by leveraging this information, businesses can make informed decisions that can help them stay competitive and succeed in today's economy.

Conclusion

Through rigorous examination of data gathered from the Bureau of Labor and Statistics, the National Center for Education Statistics, and the US Census Bureau, we conclude that there is a positive correlation between educational attainment and average income over time. More specifically, individuals with higher levels of educational attainment tend to earn more than individuals with less education.

However, correlation does not imply causation. Social science is messy and not everything is quantifiable. Immediate improvements to our model can be made by including a wider range of data from the Current Population Survey. For our model, we

only considered the years 2015 through 2021. By expanding the number of observations recorded, we can perhaps find more meaningful or durable trends in the data. Further improvements may be made by introducing datasets featuring a higher degree of granularity. The whole of the data consumed by our model has been aggregated both at the state and national level. While this kind of data is useful in finding broader trends within a population, by introducing raw data at an individual level, our model can make more meaningful predictions.

The independent variables which determine income and employment are innumerable and are likely unknowable. However, as with most other things in life, we can strive to approximate and understand these relationships by carefully interrogating the data we have available on the subject.

References

CPS Tables : U.S. Bureau of Labor Statistics. (2023, January 25). https://www.bls.gov/cps/tables.htm

Median household income, by state: Selected years, 1990 through 2019. (n.d.). https://nces.ed.gov/programs/digest/d21/tables/dt21_102.30.asp

Multi-Page Apps and URL Support | Dash for Python Documentation | Plotly. (n.d.). https://dash.plotly.com/urls

Styling. (n.d.). https://plotly.com/python/styling-plotly-express/

US Census Bureau. (2022, July 4). *Educational Attainment*. Census.gov. https://www.census.gov/topics/education/educational-attainment.html

Wan, M. (2021, December 14). *Beginner's Guide to Building a Multi-Page App using Dash, Plotly and Bootstrap*. Medium. https://towardsdatascience.com/beginners-guide-to-building-a-multi-page-dashboard-using-dash-5d06dbfc7599