Pragati Pant

Data Analytics
Level: 4000

Assignment 6: Understanding and Analyzing Adult Education Levels

Abstract and Introduction

Motivation

A nation's strength lies in its people. More specifically, it is important to analyze education as an indicator of a nation's growth. One area of education research is adult education. In this context, adult education refers to the concept that adults, approximately between the ages of 25 and 64, have finished some degree of schooling. There are three general levels of adult education: below secondary, upper secondary, and tertiary. The below secondary category includes adults who have attended some lower secondary schooling. The upper secondary category includes adults who have finished secondary schooling or high school. Finally, the tertiary category includes adults who have finished the highest degree of schooling, which entails some type of college degree. This research project aims to understand how adult education trends have changed within the United States and the world as a whole. Perhaps, this analysis will help in making predictions of where education is headed in the future.

Relevant References

Along with this research project, there are multiple studies that have been conducted about adult education around the world. These research papers all take adult education and look into a specific aspect of its effect on human life. A first study from Finland looks into how "family socioeconomic circumstances directly influence adult education level" (3). This relationship was tested using survey data from 1985 to 1995. There were over 41 thousand surveyees aged 12 to 18 years. This data was "linked with three-generation registry data of Statistics Finland until 2009" (3). Aftering analyzing the results, they found that "socioeconomic circumstances of parents and grandparents predicted adult education level. School achievement and reserve capacity dimensions of perceived health, health-promoting behaviour and social support in adolescence also positively predicted adult education (3). Ultimately, the study hoped their findings would create more fair educational opportunities for "adolescents from disadvantaged socioeconomic backgrounds" (3).

A second study "focuses on adults' levels of education-related epistemic freedom and epistemic violence" (4). A correlational research model was used to describe this problem, and the sample size of participants included 129 adults aged 22 to 67 years. Results concluded "that adults' level of accepting education-related epistemic violence and resorting to education-related epistemic violence were 'moderate' and 'low', respectively, while their enjoyment of epistemic freedom in the past was between 'moderate' and 'high' and their tendency to education-related epistemic freedom was 'high.' (4)" There was also a weak relationship between those adults who enjoyed "education-related epistemic freedom in the past" and those who resorted to epistemic violence now (4). Ultimately, the study suggested that "individuals should be provided with a freedom-based education and setting" (4).

A third study looks into a “cognitive assessment of older adults who are either illiterate or with low levels of education” (5). The goal of the study was to “review the literature on cognitive assessment tools for screening cognitive syndromes including MCI and Alzheimer's disease (AD) in older adults with low levels of education” (5). Multiple databases were searched to check for proper assessments tools; the results showed that 44 tools were available. However, “only a few of them showed diagnostic accuracy” for lower levels of educated adults (5). Ultimately, this study supports “the importance of developing specific tools for the assessment of older adults with low levels of education” (5).

Questions and Hypotheses

There are a couple questions driving this analysis. First of all, **how have adult education trends in the world (average) changed throughout the years?** More specifically, **how have adult education trends changed in the United States? How has gender played a role in the change in trends of adult education levels between the years of 1981 and 2018?** With the advancements in technology, I hypothesize adult education levels have grown positively in the world and in the United States. This means more adults are educated and belong to the tertiary education level now than before. I also hypothesize there is a growing positive correlation between years and the tertiary education level, which means more adults will be in the highest education level category in the future. Finally, I hypothesize that female education level trends have grown more positively than male education level trends in both the United States as well as in the world.

Data Description

Dataset Overview

For the purposes of analysis, the dataset used in this project is from the Organisation for Economic Co-operation and Development (OECD). This dataset has annual data points from various countries for three different levels of adult education. The three levels are defined as below secondary, upper secondary, and tertiary. All data points range from 1981 to 2018, but not every country is accounted for each year. Additionally, only data from eight countries is available for 1981. This changes as more years are included in the analysis. Therefore, in 2018, data for many more countries is available. With the addition of each new year, more and more data is available. This positive change in data availability represents the growth of the OECD in more countries.

Attributes Overview

There are eight total columns within this dataset. However, the eighth column—labelled Flag.Codes—is empty for all rows.

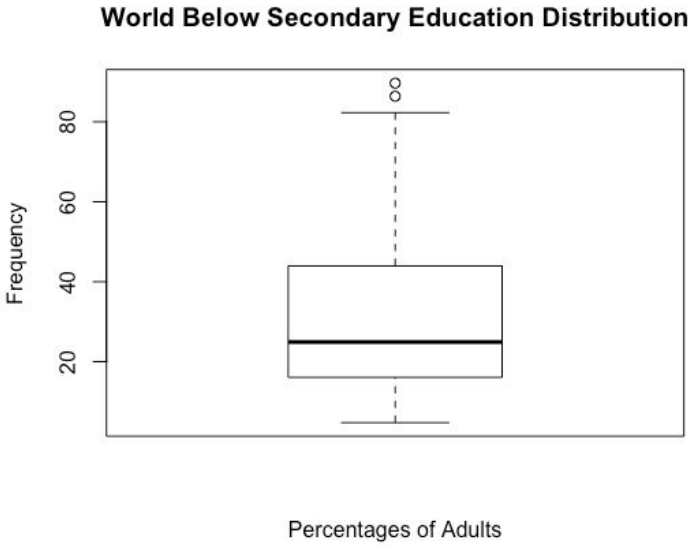
1. **Location:** The first column is LOCATION, which identifies which country the data point is from. Widely known and accepted abbreviations are used to define each country. For example, AUS refers to Australia.
2. **Indicator:** The second column is INDICATOR, which has “EDUADULT” as the entry for each row. This value serves as the reminder that the dataset is about adult education.
3. **Education Level:** The third column is SUBJECT, where the three levels of adult education are defined for each data point. The options for this column are BUPPSRY (below secondary education level), UPPSRY (upper secondary education level), UPPSRY_MEN (upper secondary education level for men), UPPSRY_WOMEN (upper secondary education level for women), TRY (tertiary education level), TRY_MEN (tertiary education level for men), TRY_WOMEN (tertiary education level for women). The percentages for tertiary and upper secondary levels are also broken up by gender, but the gender based data for below secondary level is not available.
4. **Measure:** The fourth column is INDICATOR, which has “PC_25_64” as the entry for each row. This value provides the information that the percentages available in this dataset are for adults aged 25 to 64.
5. **Frequency:** The fifth column is FREQUENCY, which has “A” as the entry for each row. This value serves as the reminder that the data points were recorded annually.
6. **Year:** The sixth column is TIME, where the year the data entry was recorded is listed. The values for this column range from 1981 to 2018.
7. **Percentage:** The seventh column is Value, which has percentages of adults. For one row, this is the percentage of adults aged 25 to 64 that fall in the category for that specific education level.

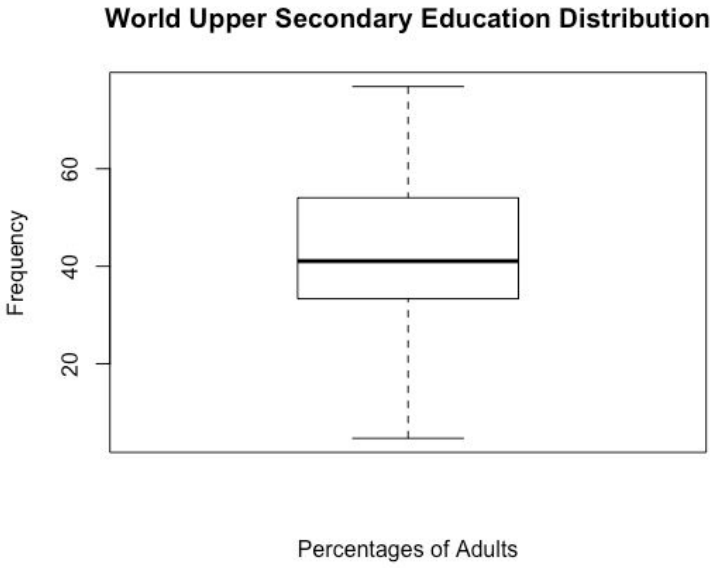
Analysis

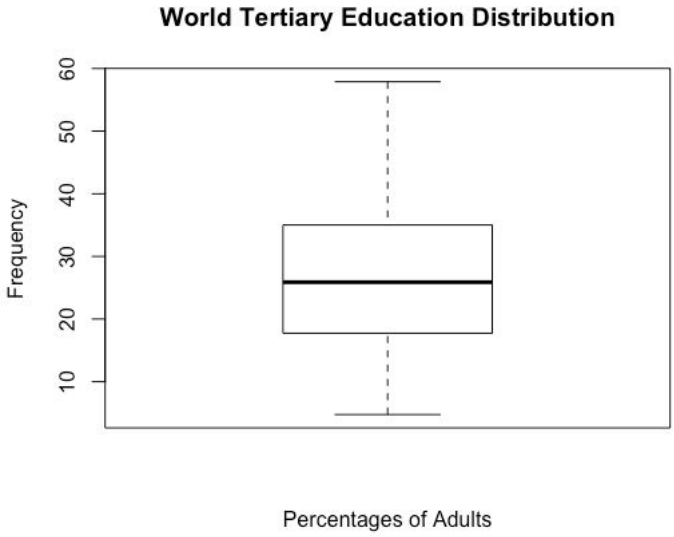
Distributions

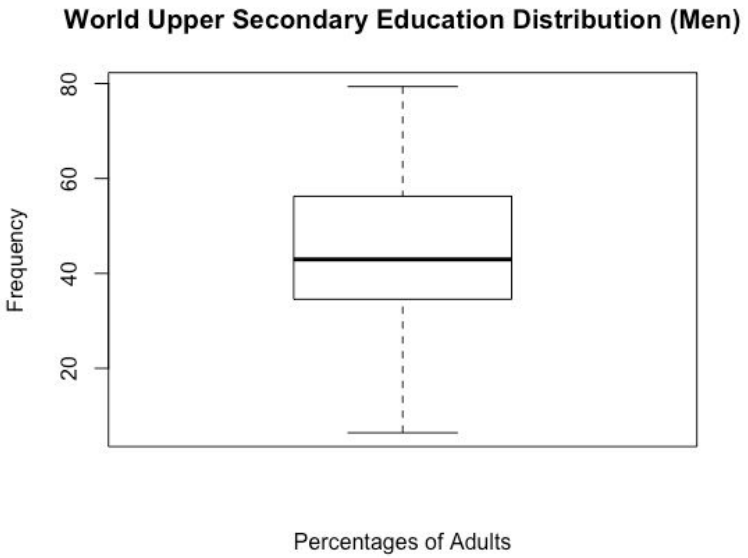
There are two main numerical columns within this dataset. The first column is TIME, which has the years each of the rows was recorded in. The second column is Value, which is the important one to plot and check for distributions.

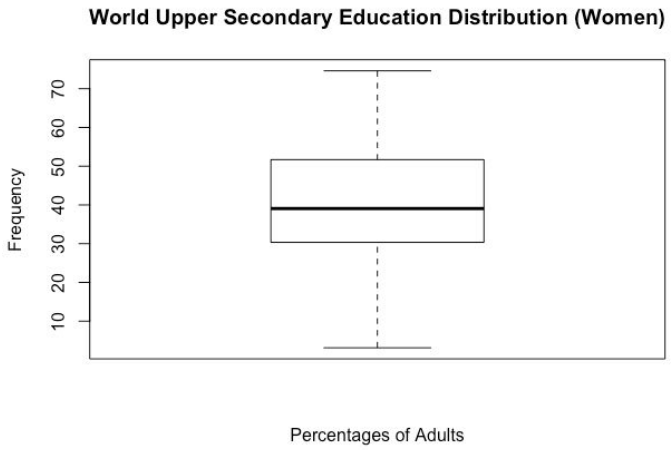
Figures 1 through 7 are boxplots for world data broken down by education levels and gender.

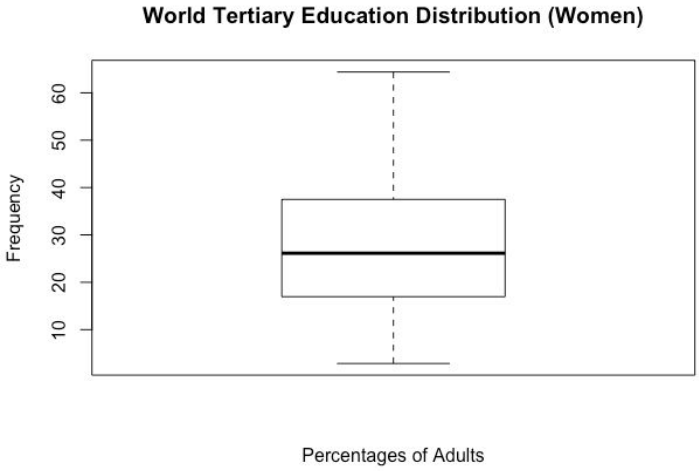


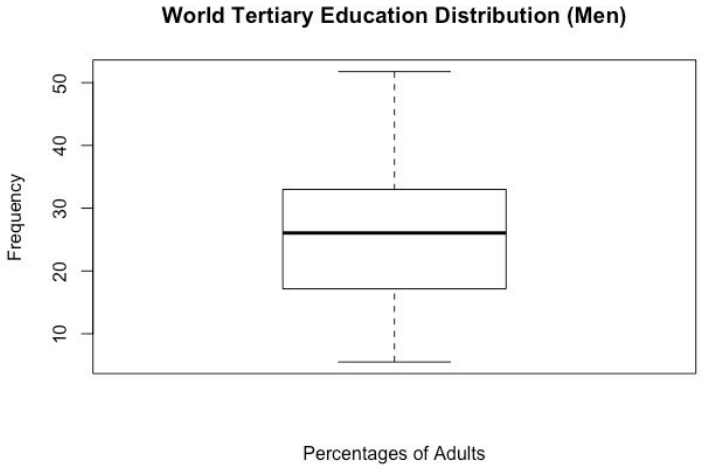




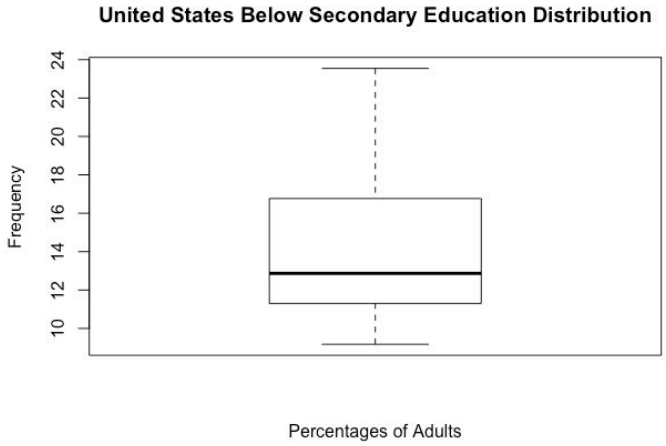


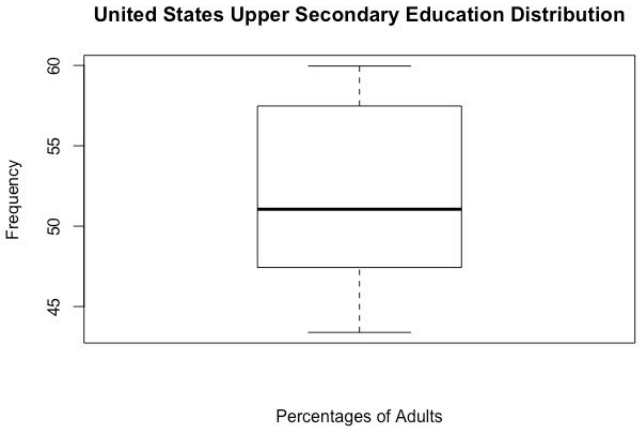


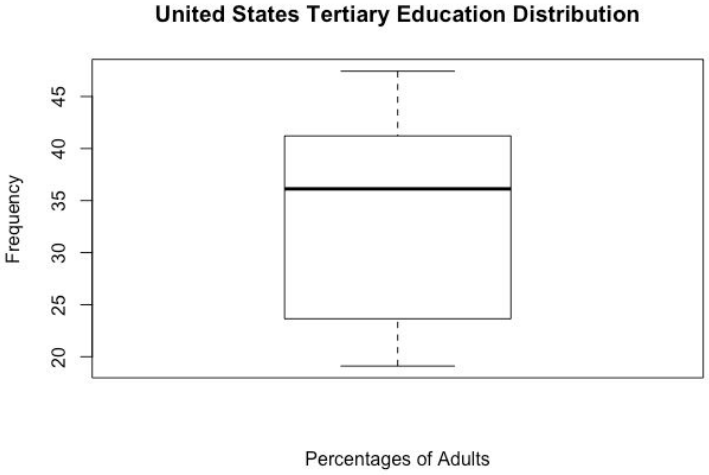


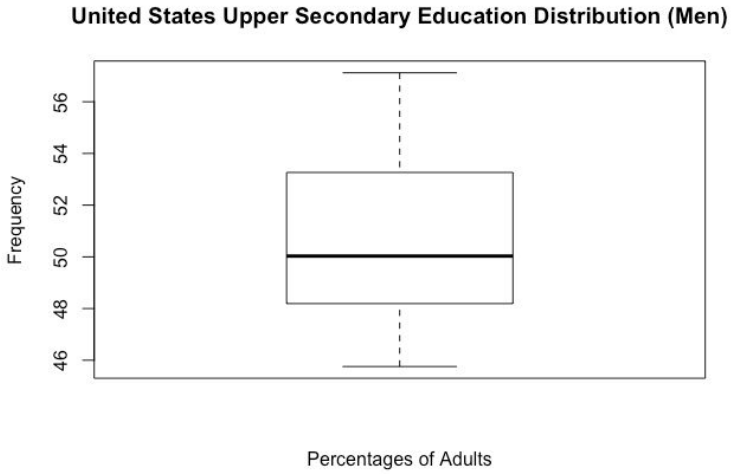


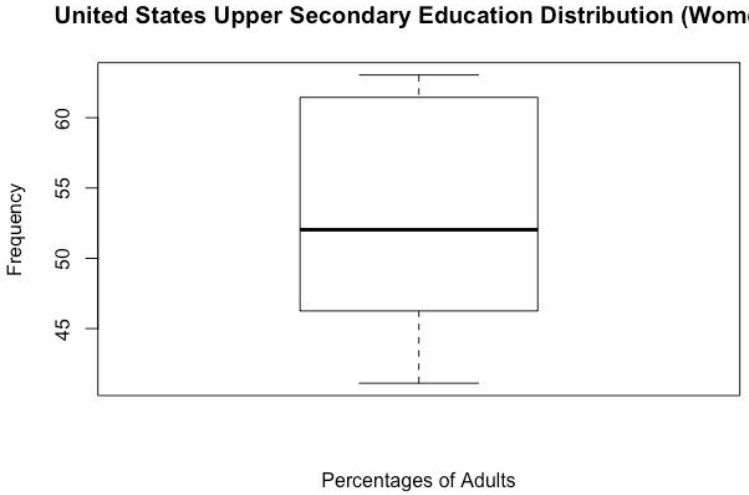
Figures 1 through 7 above show that most of the distributions do not have outliers. Only the distribution for the below secondary education level for world data (Figure 1) has two outliers. Below are the boxplots for specifically United States data.

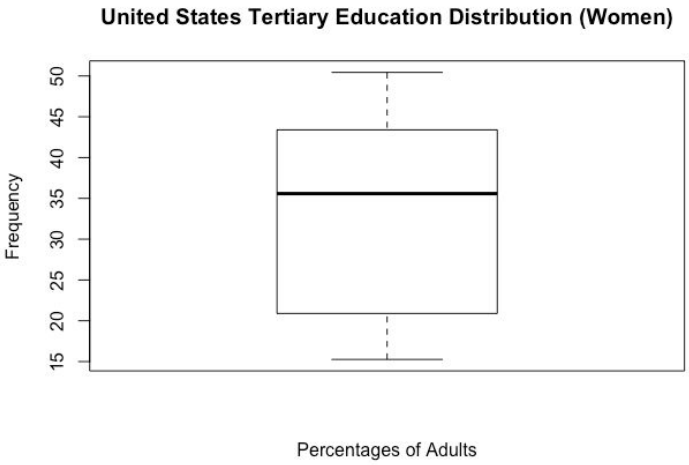

United States Below Secondary Education Distribution
Frequency
10
12
14
16
18
20
22
24
Percentages of Adults

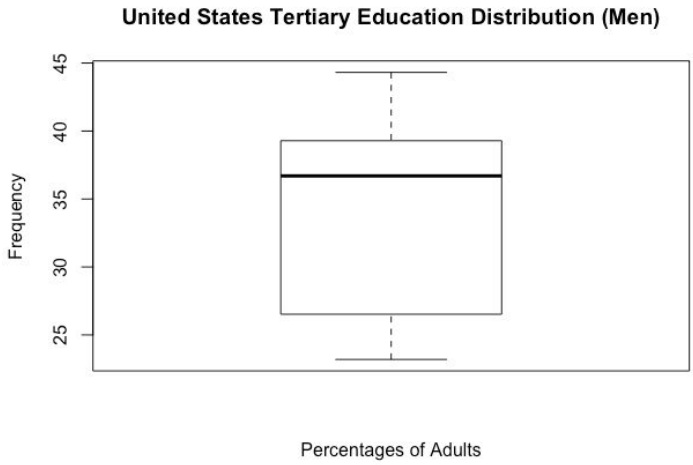

United States Upper Secondary Education Distribution
Frequency
45
50
55
60
Percentages of Adults


United States Tertiary Education Distribution
Frequency
20
25
30
35
40
45
Percentages of Adults

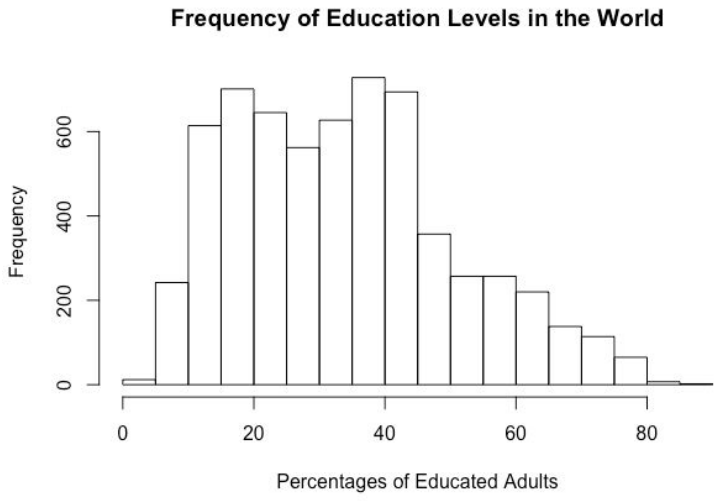

United States Upper Secondary Education Distribution (Men)
Frequency
46
48
50
52
54
56
Percentages of Adults

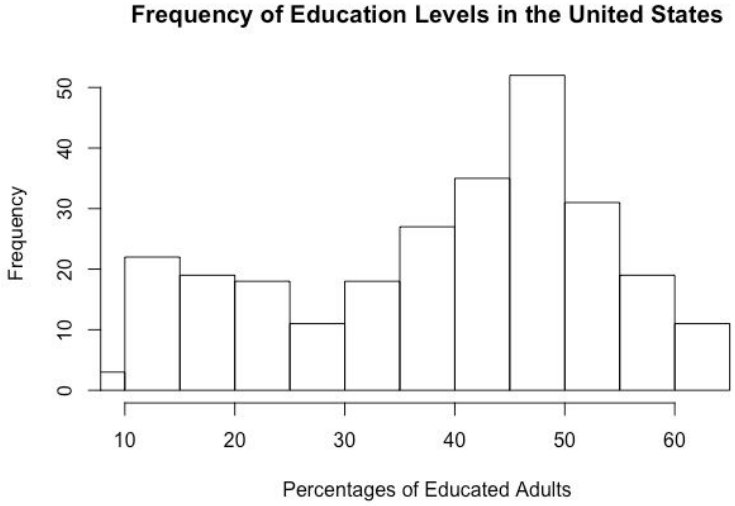

United States Upper Secondary Education Distribution (Women)
Frequency
45
50
55
60
Percentages of Adults

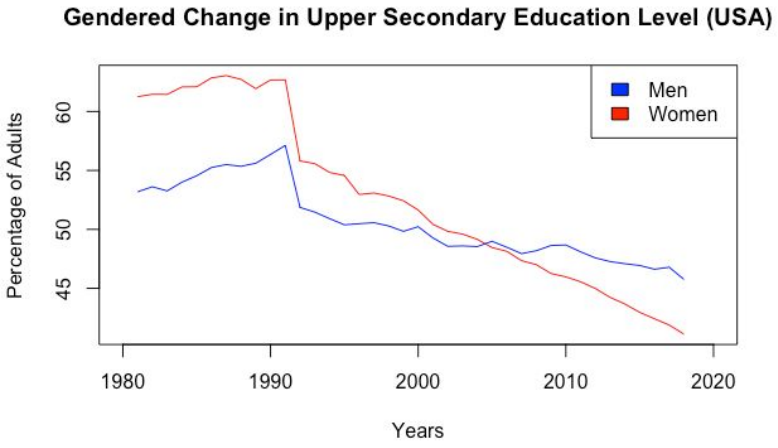


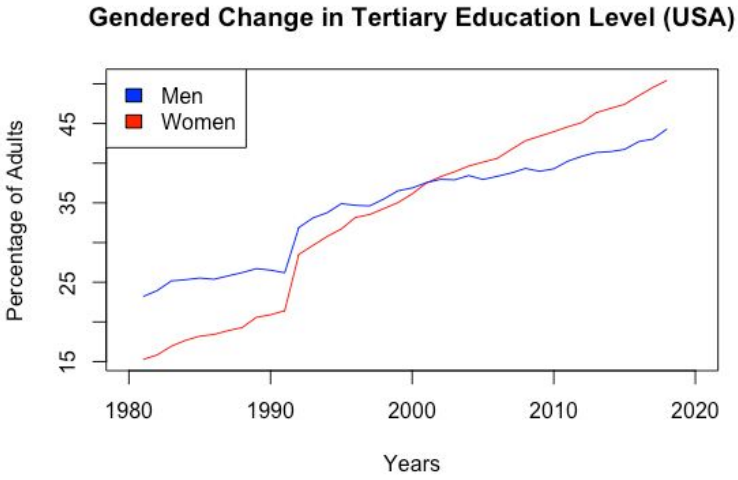


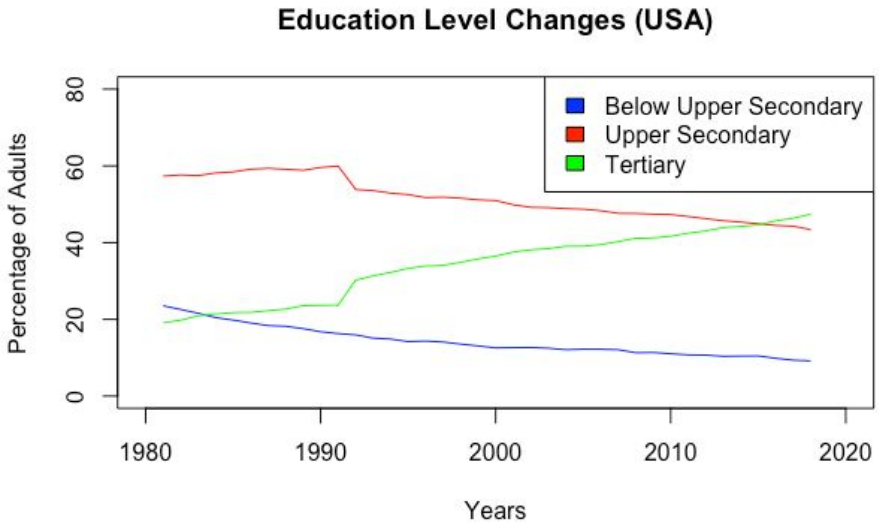
Figures 8 through 14 above show that there are no outliers for the United States data points. Figures 15 and 16 below are the histograms for both world and United States data. The plots show that there are not normal distributions for either of the filtered data points; the distributions are either skewed or have multiple peaks.

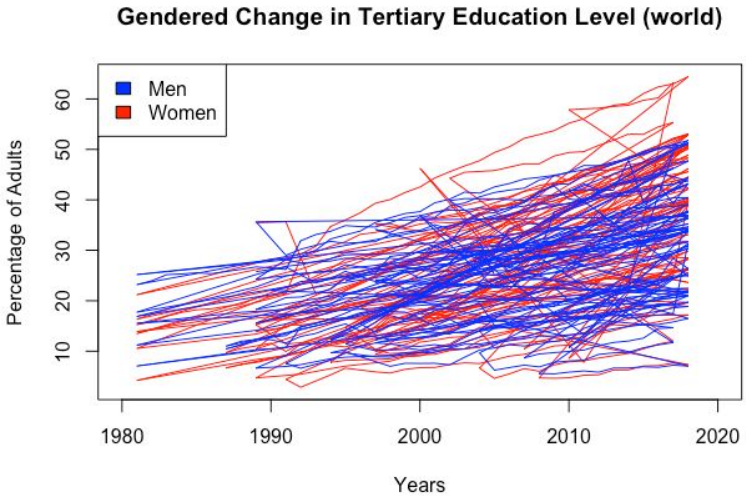


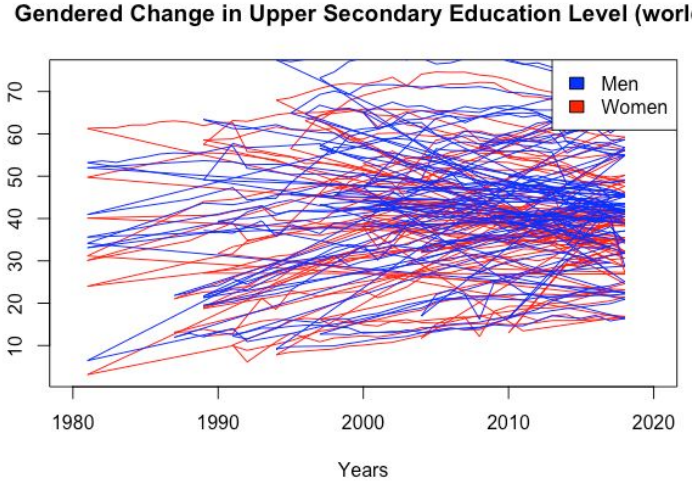


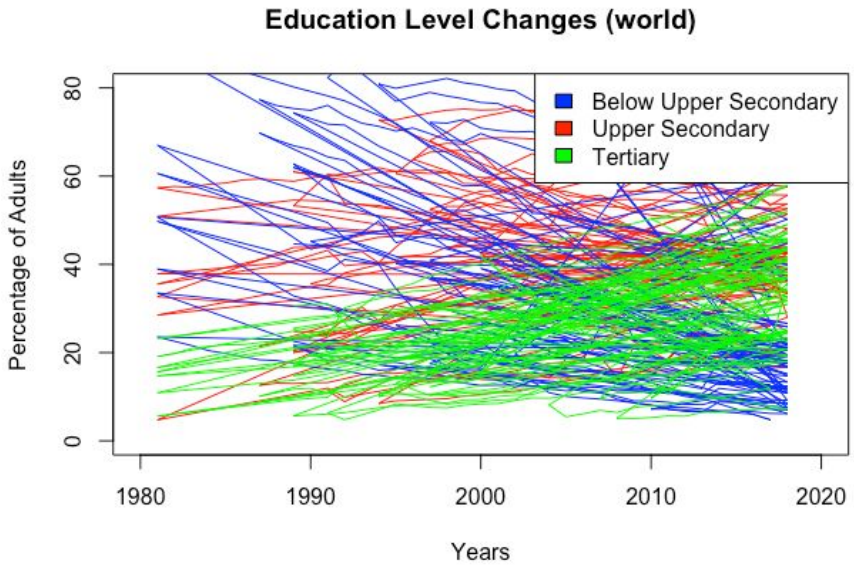

Gendered Change in Upper Secondary Education Level (USA)
Men
Women
Percentage of Adults
45
50
55
60
1980
1990
2000
2010
2020
Years


Gendered Change in Tertiary Education Level (USA)
Men
Women
Percentage of Adults
15
25
35
45
1980
1990
2000
2010
2020
Years


Education Level Changes (USA)
Below Upper Secondary
Upper Secondary
Tertiary
Percentage of Adults
0
20
40
60
80
1980
1990
2000
2010
2020
Years


Gendered Change in Tertiary Education Level (world)
Men
Women
Percentage of Adults
10
20
30
40
50
60
1980
1990
2000
2010
2020
Years


Gendered Change in Upper Secondary Education Level (world)
Men
Women
10
20
30
40
50
60
70
1980
1990
2000
2010
2020
Years



Figures 17 through 22 above are very simple, overlaying graphs of the changing adult education levels in the United States and in the world. Better plots will be needed to understand the linear regression model.

Uncertainty and Bias

While metadata and information is provided by the OECD, there are still a few areas of uncertainty and bias. Firstly, gender is binary in the context of this dataset. This is not inclusive of non-binary individuals, and therefore is not predictive of all people in the world and in the United States. Secondly, there is no strict definition on the OECD's website of what below secondary, upper secondary, and tertiary levels are. Even using other sources, the levels' definitions are loosely defined and accepted. Finally, not every country's adult education information is accounted for each year. This makes it really hard to accurately measure and make predictions off of the linear regression model for the world. For this reason, the world data has been filtered to see if the linear relationship is better for the United States.

Model Development and Application

Overview

A regression model will be used to describe the data. To gain full understanding of the dataset, multiple variations of this regression model will be needed. The main variables for the model are percentage of adults educated and year for each row. Additionally, the data will need to be filtered by country and also by education level. As stated before, there are three main options for adult education level; two of the levels have data broken up by gender which will be assessed too. The important columns for this model are Value, TIME, LOCATION, and SUBJECT. Ultimately, the model will be scaled to look at the changes in adult education levels within the United States as well as within the entire world. The goal of the regression model is to show and predict the relationship between how much of the population is educated and time.

Model Using World Data

The first attempt of this regression model uses all data points from countries around the world. Except for any unknown country types or wrongful rows, all data points are included. All plots feature the different countries' data using a color based legend.

Figure 23 below shows the regression model for all countries between the years of 1981 and 2018 for the below upper secondary education level. The adjusted R-squared value for this plot is 0.09864, which means there is a very weak linear relationship between adults in the below upper secondary education level and time.

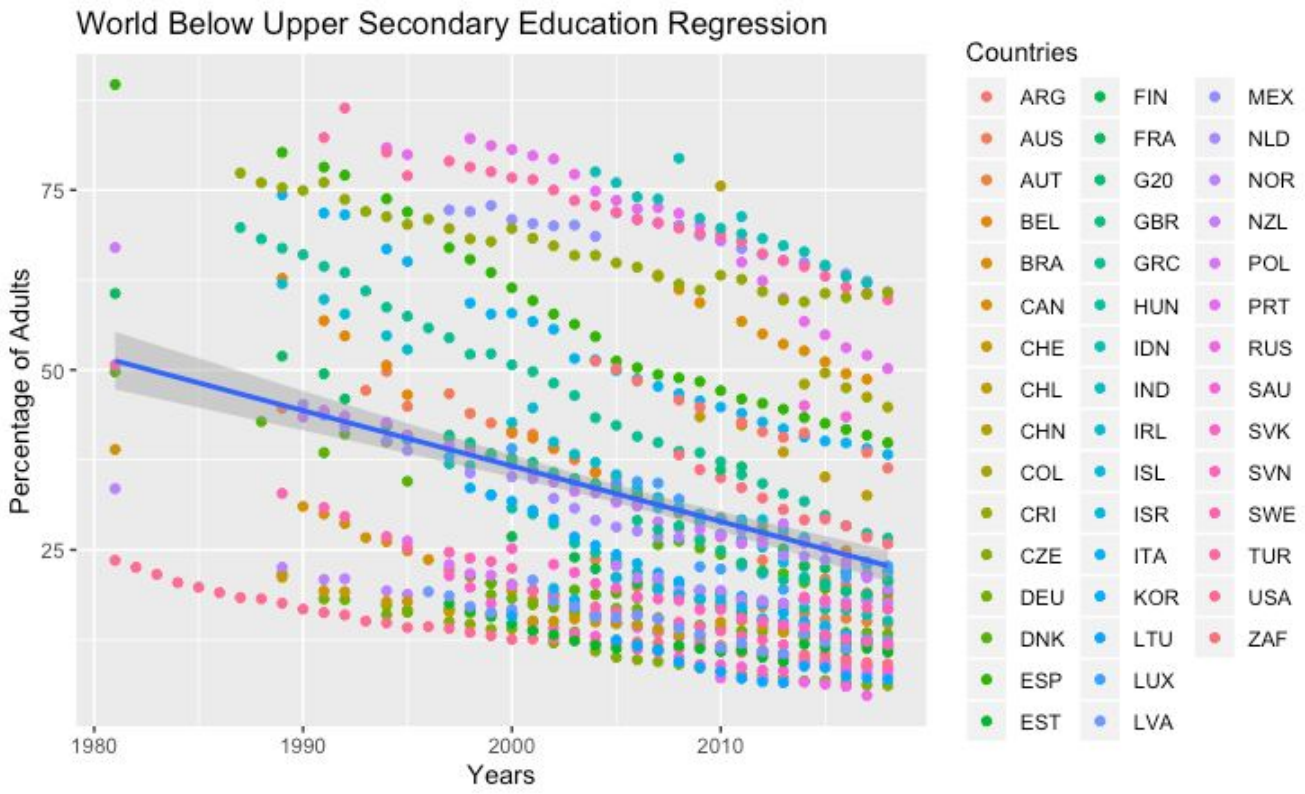


Figure 24 plot below shows the regression model for all countries between the years of 1981 and 2018 for the upper secondary education level. The adjusted R-squared value for this plot is 0.0049, which again means there is a very weak linear relationship between adults in the upper secondary education level and time.

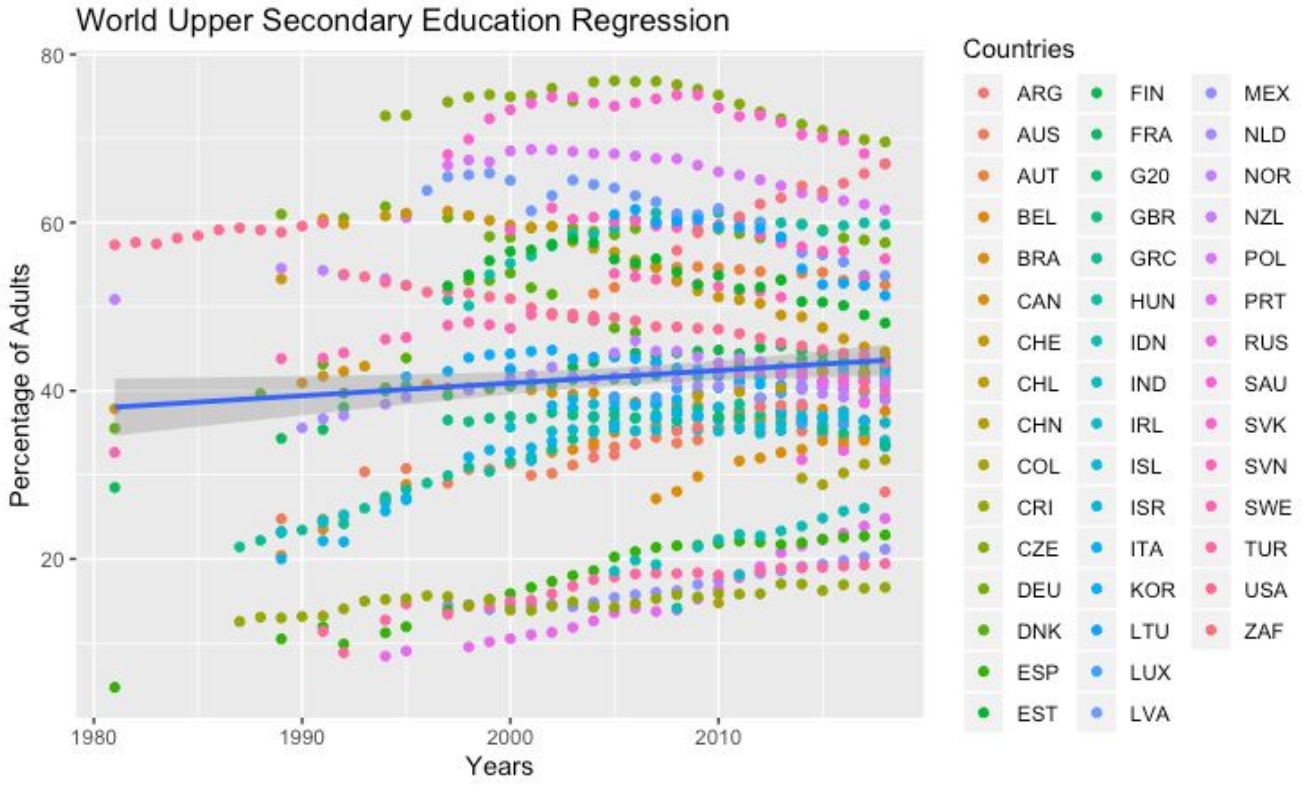
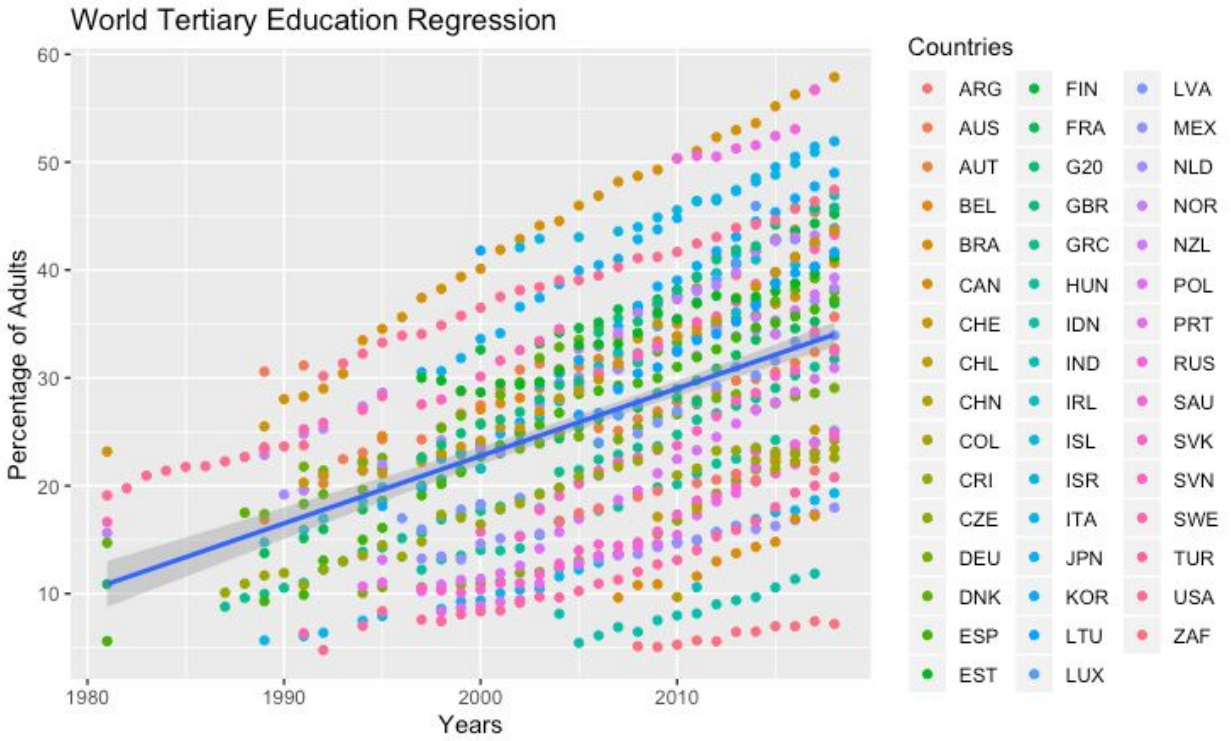


Figure 25 plot below—the final plot for this section—shows the regression model for all countries between the years of 1981 and 2018 for the tertiary education level. The adjusted R-squared value is 0.2054, which still represents a fairly weak linear relationship between adults in the tertiary education level and time. While the adjusted R-squared value for this model is better than the adjusted R-squared values for both below upper secondary and upper secondary levels, the linear relationship is still very weak.

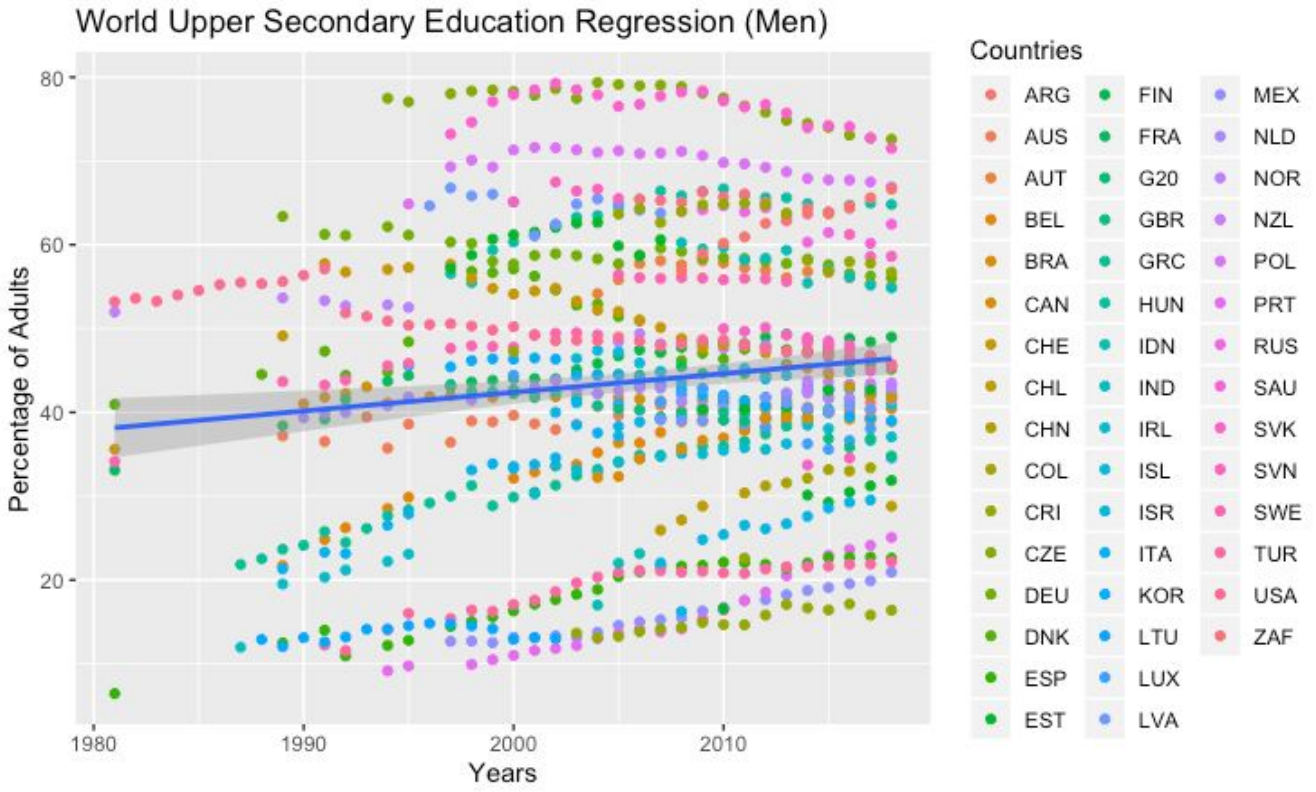


Model Using World Data (Gender)

The second attempt of this regression model uses all data points from countries around the world, but filters the data based on gender. Except for any unknown country types or wrongful rows, all data points are included. All plots feature the different countries' data using a color based legend.

Figure 26 below shows the regression model for all countries between the years of 1981 and 2018 for men in the upper secondary education level. The adjusted R-squared value for this plot is 0.01102, which means there is a very weak linear relationship between men in the below upper secondary education level and time.

Figure 27 also below shows the regression model for all countries between the years of 1981 and 2018 for men in the tertiary education level. The adjusted R-squared value for this plot is 0.1164, which is better than the adjusted R-squared for the first plot; regardless, there is still a very weak linear relationship between men in the tertiary education level and time.


World Upper Secondary Education Regression (Men)
Percentage of Adults
80
60
40
20
1980
1990
2000
2010
Years
Countries
ARG
AUS
AUT
BEL
BRA
CAN
CHE
CHL
CHN
COL
CRI
CZE
DEU
DNK
ESP
EST
FIN
FRA
G20
GBR
GRC
HUN
IDN
IND
IRL
ISL
ISR
ITA
KOR
LTU
LUX
LVA
MEX
NLD
NOR
NZL
POL
PRT
RUS
SAU
SVK
SVN
SWE
TUR
USA
ZAF

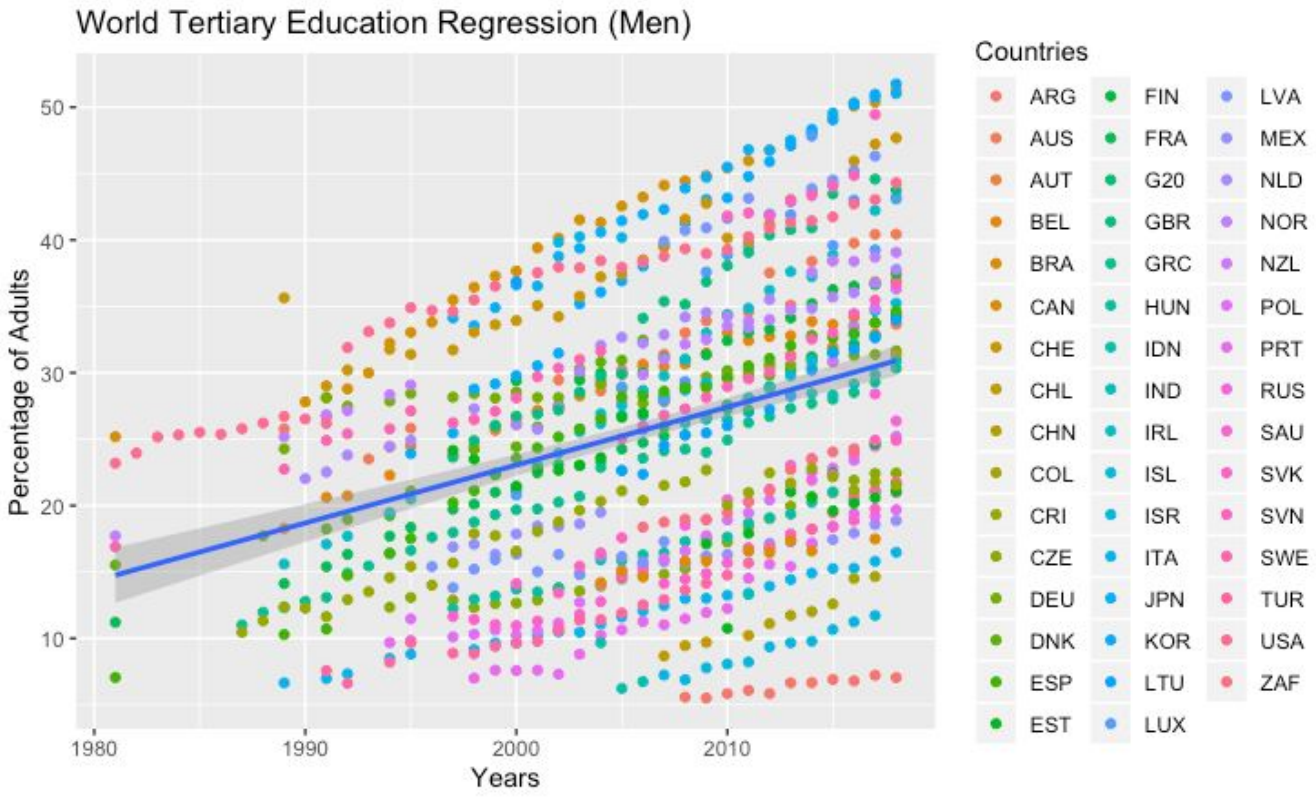

World Tertiary Education Regression (Men)
Percentage of Adults
50
40
30
20
10
1980
1990
2000
2010
Years
Countries
ARG
AUS
AUT
BEL
BRA
CAN
CHE
CHL
CHN
COL
CRI
CZE
DEU
DNK
ESP
EST
FIN
FRA
G20
GBR
GRC
HUN
IDN
IND
IRL
ISL
ISR
ITA
JPN
KOR
LTU
LUX
LVA
MEX
NLD
NOR
NZL
POL
PRT
RUS
SAU
SVK
SVN
SWE
TUR
USA
ZAF

Figure 28 below shows the regression model for all countries between the years of 1981 and 2018 for women in the upper secondary education level. The adjusted R-squared value for this plot is 0.0006192, which means there is a very weak linear relationship between women in the below upper secondary education level and time.

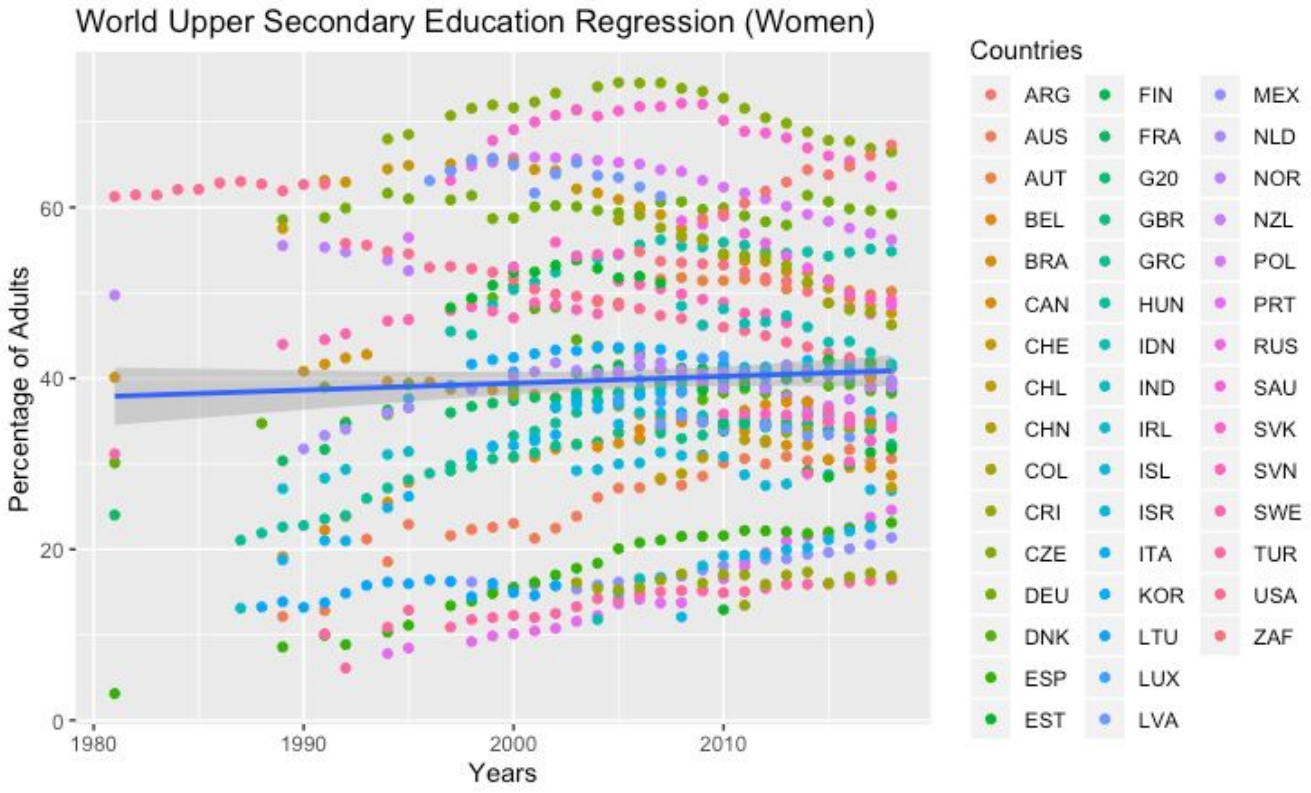
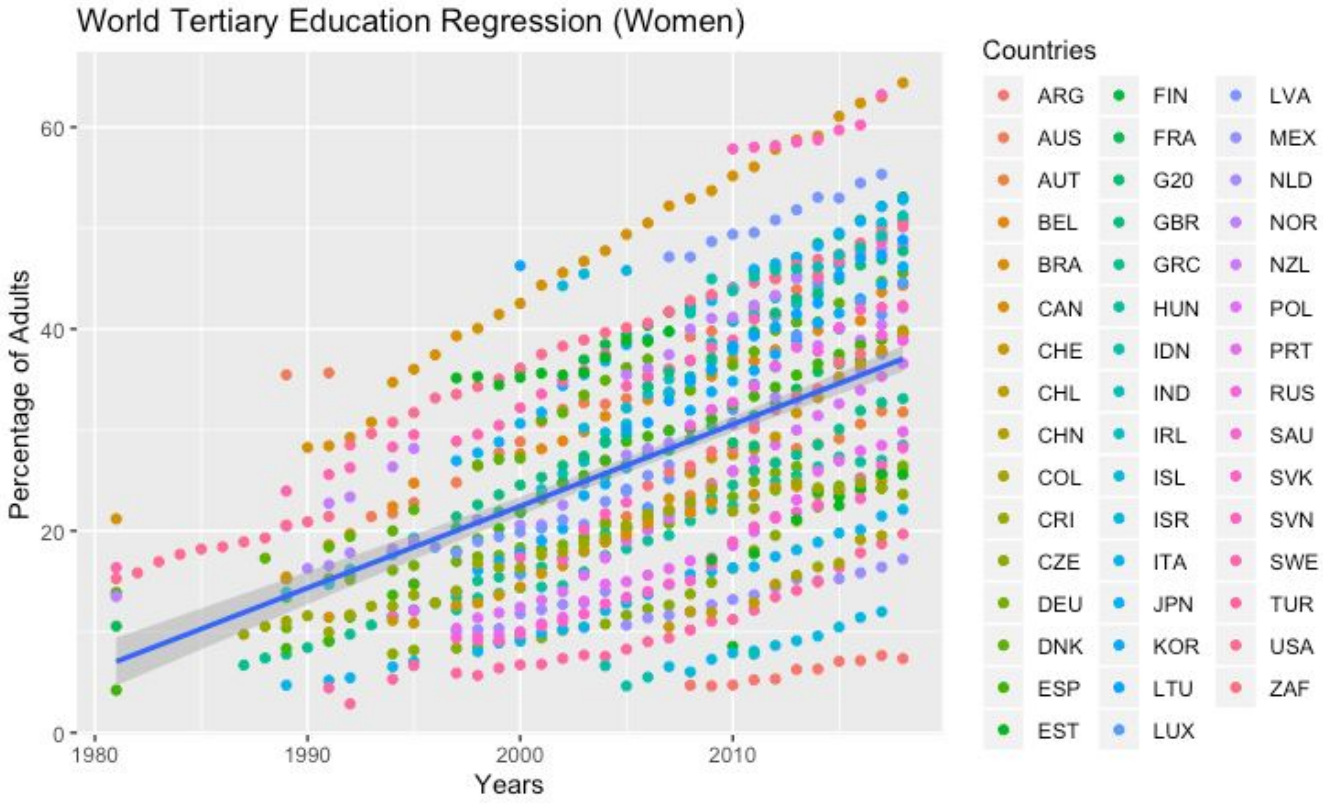


Figure 29 below shows the regression model for all countries between the years of 1981 and 2018 for women in the tertiary education level. The adjusted R-squared value for this plot is 0.2607, which is the best adjusted R-squared for the gender based plot. However, the R-squared value is still very far from 1, which means there is still a very weak linear relationship between women in the tertiary education level and time.



Model Using United States Data

The third attempt of this regression model uses data points just from the United States. The goal of plotting and modeling this filtered United States data is to see if narrowing the scope results in a better linear correlation. All plots have data points that range from 1981 to 2018.

Figure 30 below shows the regression model for the United States between the years of 1981 and 2018 for the below upper secondary education level. The adjusted R-squared value for this plot is 0.9148. This value is very close to 1, which represents a very strong negative linear correlation between American adults in the below upper secondary education level and time.

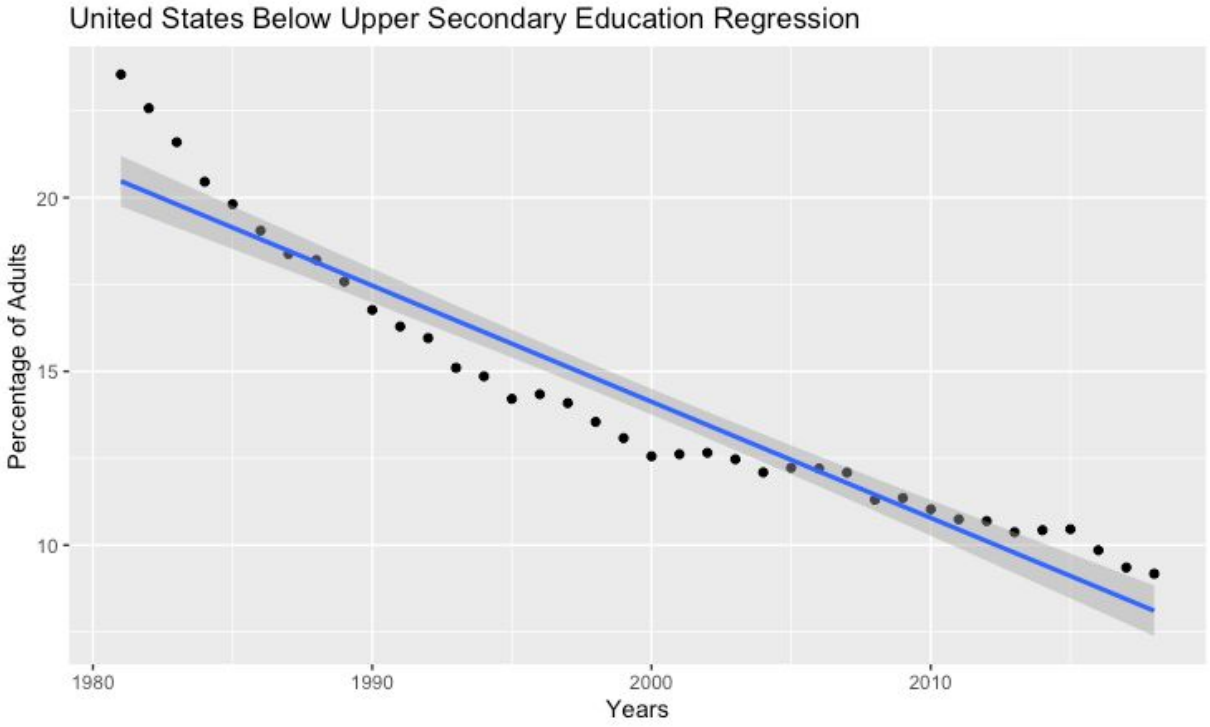
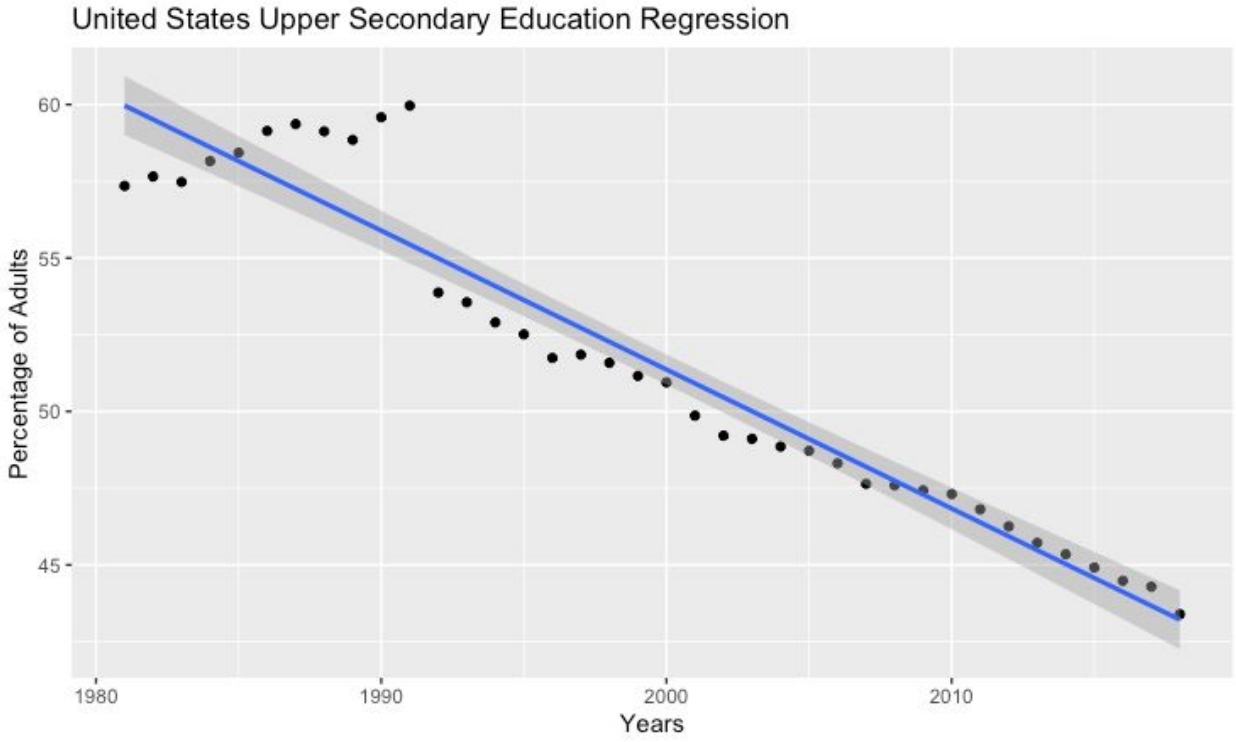


Figure 31 below shows the regression model for the United States between the years of 1981 and 2018 for the upper secondary education level. The adjusted R-squared value for this plot is 0.9191. This value is very close to 1, which represents a very strong negative linear correlation between American adults in the upper secondary education level and time.



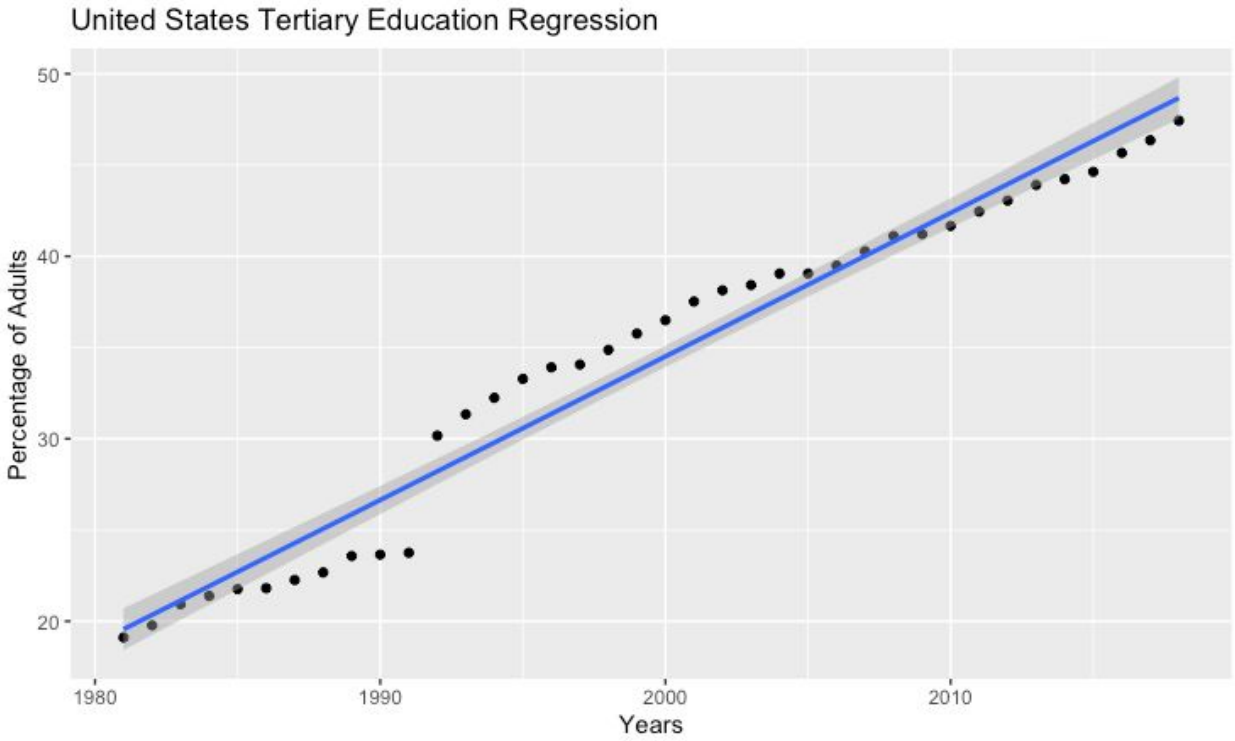


Figure 32 above shows the regression model for the United States between the years of 1981 and 2018 for the tertiary education level. The adjusted R-squared value for this plot is 0.9602. This value is very close to 1, which represents a very strong positive linear correlation between American adults in the tertiary education level and time. Of the three models for the United States, this model has the strongest supported positive correlation between the variables.

Model Using United States Data (Gender)

The third attempt of this regression model uses data points just from the United States, but filters the data based on gender. The goal of plotting and modeling this filtered United States data is to see if narrowing the scope results in a better linear correlation. All plots have data points that range from 1981 to 2018.

Figure 33 below shows the regression model for the United States between the years of 1981 and 2018 for women in the upper secondary education level. The adjusted R-squared value for this plot is 0.9491, which represents a strong negative linear correlation between American women in the upper secondary education level and time.

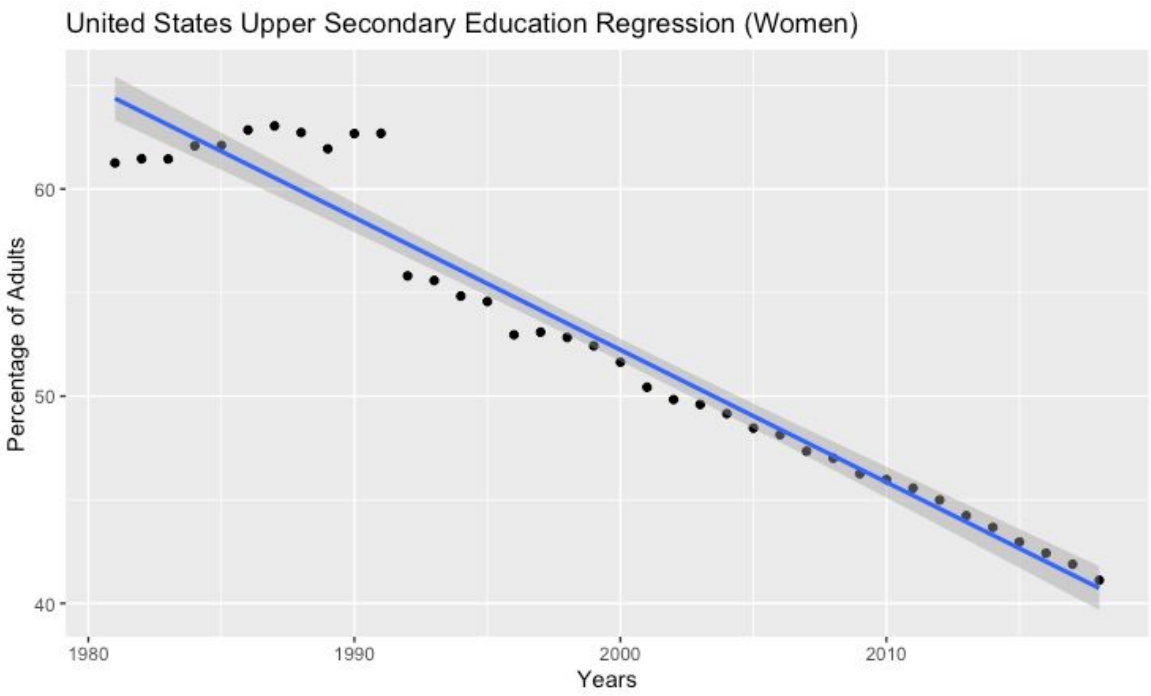


Figure 34 below shows the regression model for the United States between the years of 1981 and 2018 for women in the tertiary education level. The adjusted R-squared value for this plot is 0.971. This value is very close to 1, which represents a very strong positive linear correlation between American women in the tertiary education level and time.

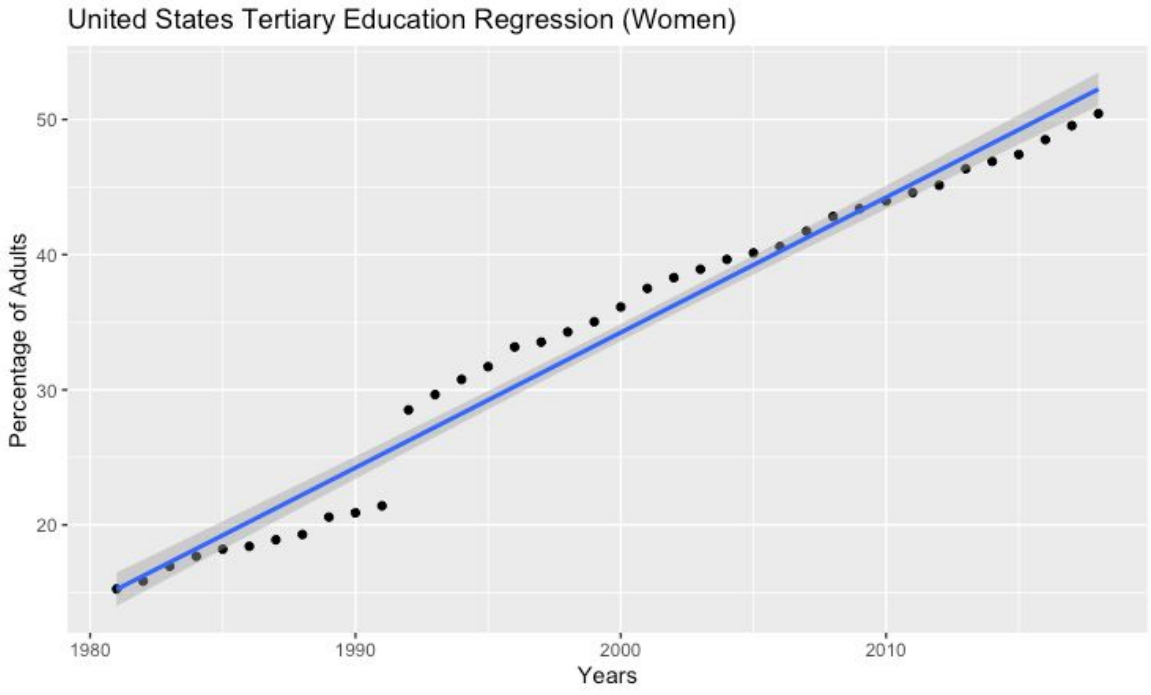


Figure 35 below shows the regression model for the United States between the years of 1981 and 2018 for men in the upper secondary education level. The adjusted R-squared value for this plot is 0.8144. This value represents a fairly strong negative linear correlation between American men in the upper secondary education level and time.

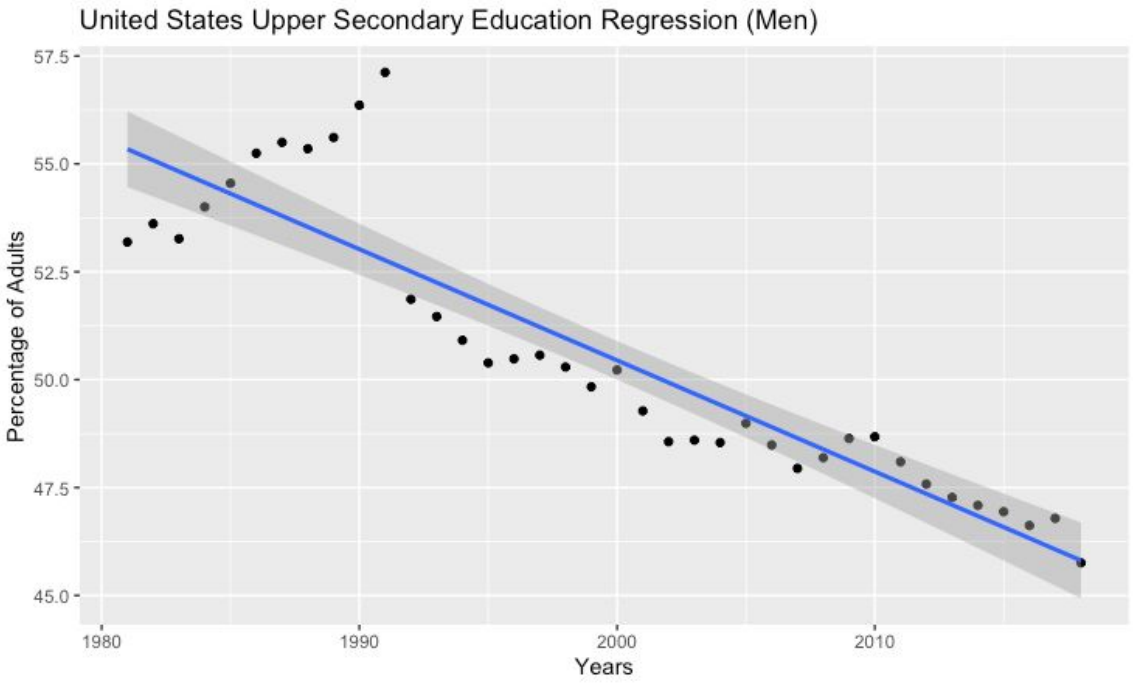
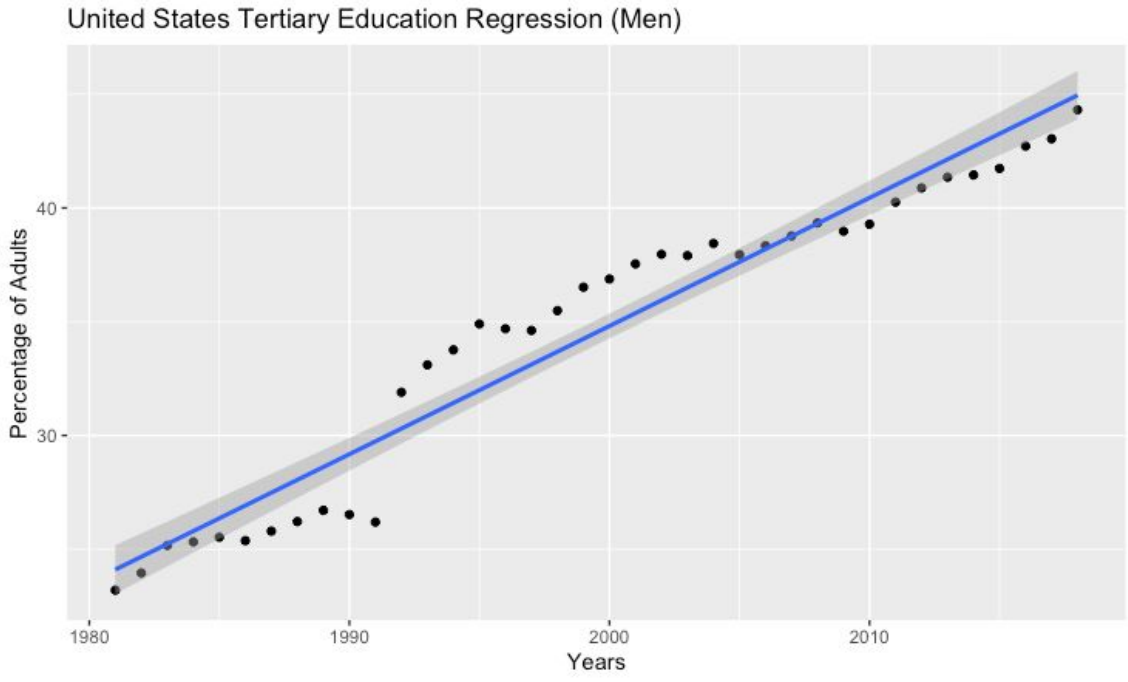


Figure 36 below shows the regression model for the United States between the years of 1981 and 2018 for men in the tertiary education level. The adjusted R-squared value for this plot is 0.9343. This value is close to 1, which represents a strong positive linear correlation between American men in the tertiary education level and time.



Modeling Conclusions

Predictions for future analysis can be made based on the multiple variations of the regression model used for this dataset. Overall, the models for all of the world data resulted in very weak adjusted R-squared values. The result was similar when the world data models were broken down by gender. When the world data was filtered to just look at data from the United States, the models—both overall and gender based ones—resulted in much more strong linear correlations. Altogether, predictions for future adult education trends in the United States can be made confidently using the United States linear regression models.

Conclusions and Discussion

Overview

The questions initially asked were addressed by the regression analysis conducted. The world regression models did not support the hypothesis that the world adult education levels trends have changed positively. The adjusted R-squared value for the world models showed very weak linear relationships between time and percentages of adults in the three education levels. On a different note, the United States regression models supported the hypothesis that adult education trends have grown positively. The adjusted R-squared values for these models were close to 1. Looking specifically, more adults are in the tertiary education level now than before, and less adults are in the below upper secondary and upper secondary levels now than before—this supports the hypothesis put forth. Additionally, it is also supported that more women are in the tertiary education level now than before. Altogether, while the hypothesis is supported for the United States, more sampling and analysis will need to be done to really form a conclusion about this. For now, the research is inconclusive for the world data, but supportive of the filtered United States data.

Changes in Process

While the original goals of the project remained the same throughout, some of the approaches changed as more information about the dataset was revealed. Initially, the idea was to use a linear regression model to predict future adult education levels' trends for countries across the world. However, as it was noticed that not every country is accounted for each year, the scope of the analysis was changed to focus more strictly for the United States. Of the many countries, this was one that had data points for all years between 1981 and 2018. While the regression model was built for the entire world, the regression model showed more promise for the United States.

Predictions

Using the United States regression models, predictions can be made about the future of adult education trends in the United States. By 2020, this model predicts that 50.24887 of American adults will belong to the tertiary or highest level of education. 42.3097 percent will belong to the upper secondary education level, and 7.441435 will belong to the below secondary level. By 2050, 73.85634 of American adults will belong to the tertiary education level, and by 2075, 93.52923 of American adults will be in the tertiary category. This positive growth in tertiary education level predictions shows that this regression model supports the original hypothesis for the United States. Predictions for the world can not be made since the world regression models have very weak linear correlation.

Future Analysis

To further this analysis, more data can be used to better understand adult education levels around the world. Perhaps, more sources can be corporated to cross-reference any previous analysis of adult education. Other countries aside from the United States can also be looked at

more specifically to get a better understanding of how they compare to the United States. Finally, a different type of model can be applied to this dataset. For example, a clustering model would allow for grouping of countries to show regional adult education growth. Ultimately, any other form of research would help in better understanding what direction adult education levels' trends are headed around the world.

References

1. Adult education level. Retrieved at https://data.oecd.org/eduatt/adult-education-level.htm
2. *EDUCATION INDICATORS: An International Perspective*. Retrieved at https://nces.ed.gov/Pubs/eiip/eiip1s01.asp
3. Acacio-Claro, P. J., Doku, D. T., Koivusilta, L. K., Rimpelä, A. H. (2018). How socioeconomic circumstances, school achievement and reserve capacity in adolescence predict adult education level: a three-generation study in Finland. International Journal of Adolescence & Youth; Sep2018, Vol. 23 Issue 3, p382-397, 16p.
4. Kartal, O. Y., Yazgan, A. D., Avci, E. (2018). An Investigation into the Relationship between Adults' Levels of Education-Related Epistemic Freedom and Epistemic Violence. International Education Studies, v11 n10 p96-104 2018. 9 pp.
5. Tavares-Júnior, J. W. L., de Souza, A. C. C., Alves, G. S., Bonfadini, J. de C., et al. (2019). Cognitive Assessment Tools for Screening Older Adults With Low Levels of Education: A Critical Review. Frontiers in Psychiatry; 12/13/2019, Vol. 10, p1-12, 12p.
6. Github code URL: https://github.com/pantp/DataAnalyticsSpring2020/tree/master/project