

Data Exploration & Preparation

Introduction to Data Analytics



May 15, 2020

Name: James want

Student Number: 98126764

**1A Initial Data Exploration**

1. **Attribute Types**

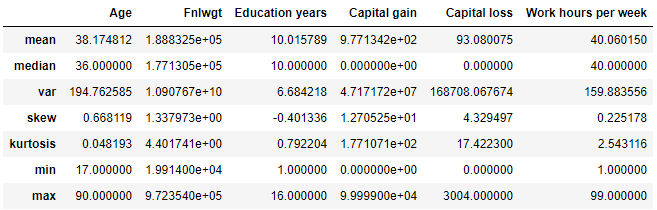
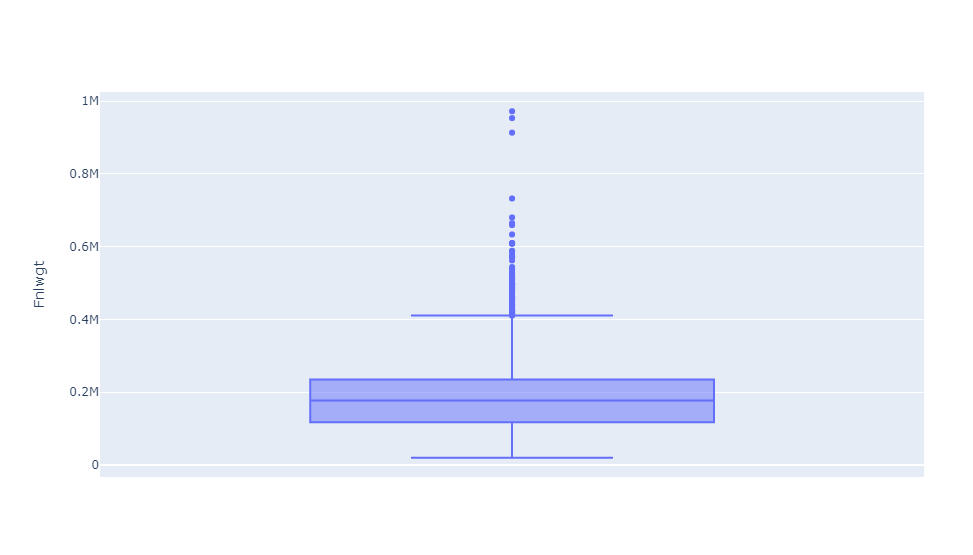
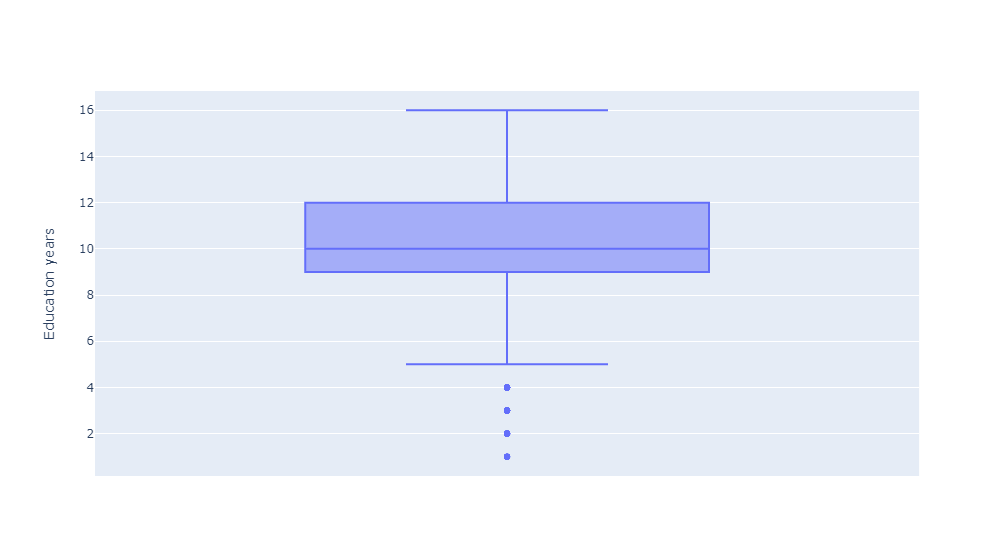
|  |  |  |
| --- | --- | --- |
| Attribute | Type | Observations |
| ID | nominal | Used to track entries for the data. Could potentially represent the order entries were gathered but there is no current evidence to confirm this, will be ignored going forward. |
| Age | ratio | Standard age values |
| Employment class | nominal | Should have some clear relationships to other factors such as education, income, and work hours. |
| Fnlwgt | ratio | With the assumption that this dataset is obtained from <https://www.kaggle.com/uciml/adult-census-income> the full definition of fnlwgt can be found there. It defines that ‘People with similar demographic characteristics should have similar weights’. So, a weight reflects the general portion of the population that can be applied to that entry. |
| Education level | nominal | nominal to start with however with some groupings can be represented as ordinal. Ie: you can easily compare year 9 < year 10 < bachelor but you can't compare Masters vs Doctorate but they can be grouped together. I think this will lend to more beneficial relationships with other attributes. For now, we will treat as nominal and transform into ordinal later. |
| Education years | ratio | can bring some useful relationships similar to education level but I think it is slightly less useful since we do not have a gage of what is the starting point and how can you compare levels of education |
| Marital status | nominal | Fairly standard marital status |
| Occupation | nominal | Allot of variety in this one which could make it difficult to create relationships and leaves a lot of questions. 'Other-services' vs ‘?’ makes it uncertain whether the entry was empty of if there is an unknown option. Is there unemployed? Are there things like hospitality workers that go under 'Other-services'. |
| Relationship status | nominal | Had to find out that own child means sons or daughter of a family. |
| Race | nominal | Fairly standard race classifications |
| Sex | nominal | Fairly standard sex classification. Can be easily binarized for different operations |
| Capital gain | ratio | Capital gains is heavily related to capital loss. I am assuming its related to how much an asset you own has increased in value |
| Capital loss | ratio | Capitol loss is how much an investment has been lost. Interesting to note that we do not have a time that this is measured over |
| Work hours per week | ratio | A measurement of time with a define zero point. I think it will relate heavily to employment class. |
| Native country | nominal | Fairly straight forward but hard to know if '?' represent missing data or an unknown choice. Majority of records seem to be from the United States |
| Salary | Ordinal | Even though income be a ratio, since we only have a defined point of less that or greater than we cannot treat this attribute as such. |

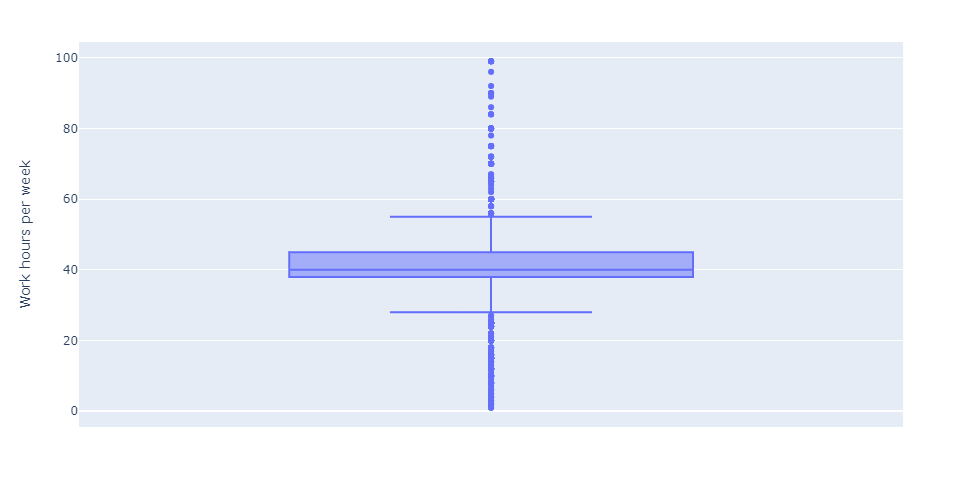
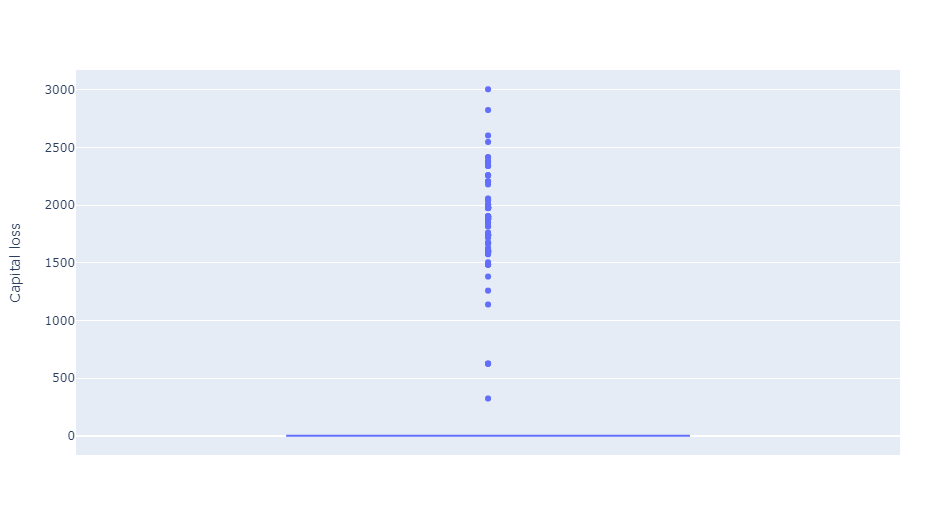
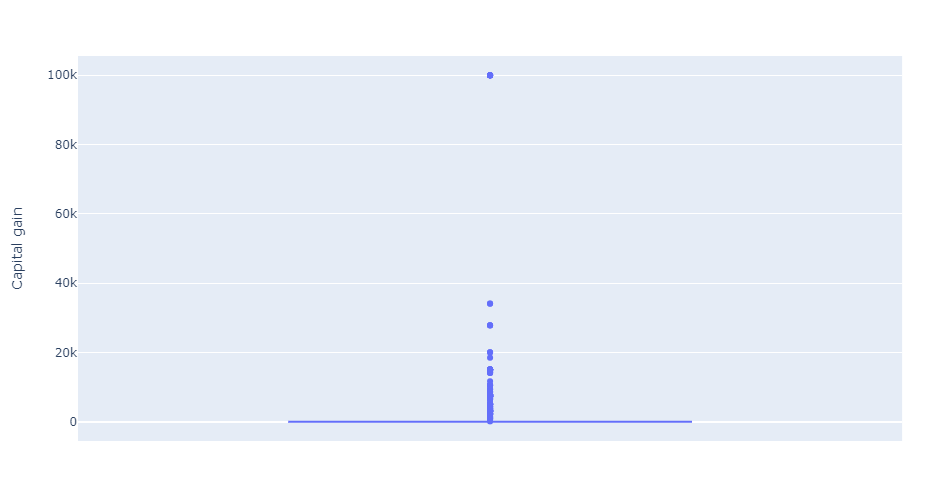
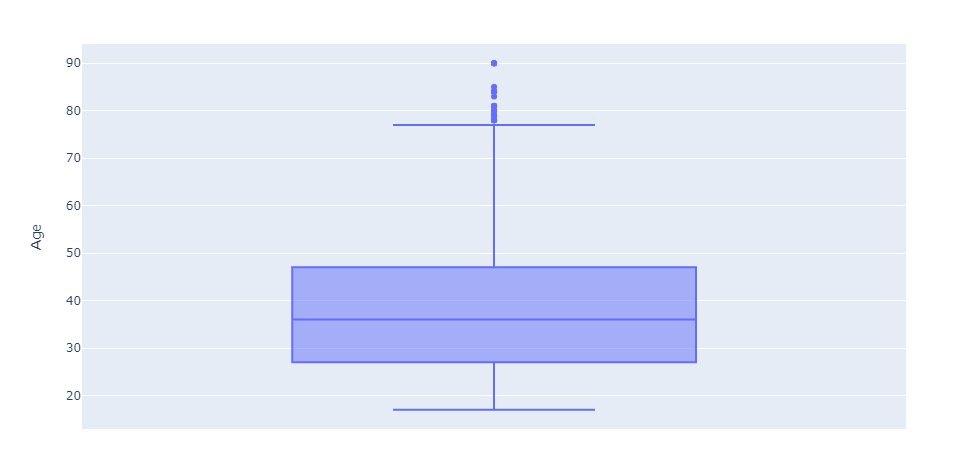
1. **Summarize Properties**

All statistics were gathered using the Pandas data frames library with no options selected.

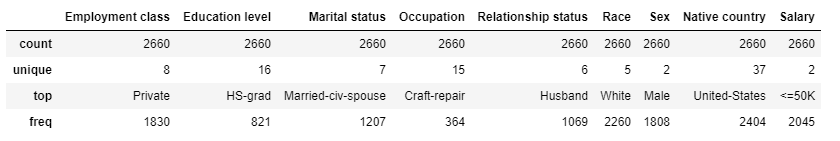
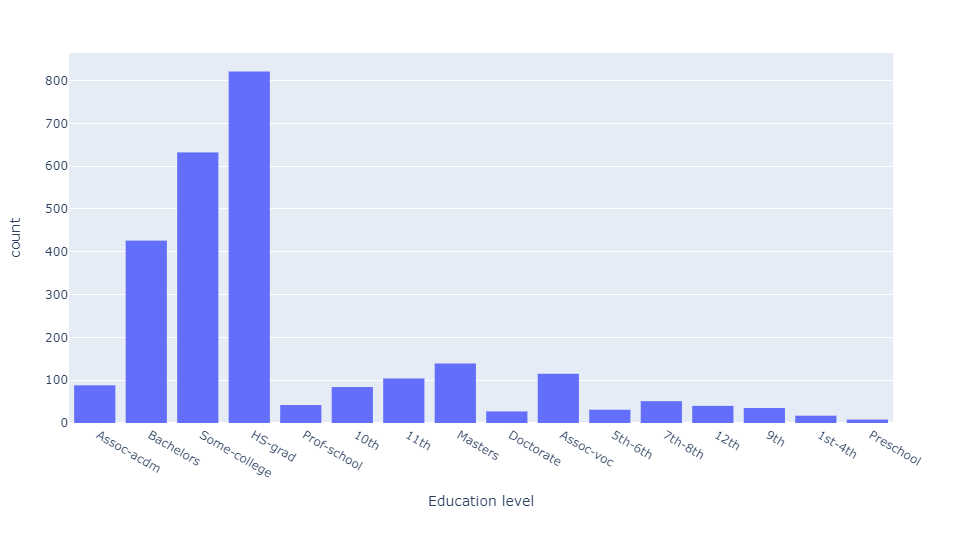
**Numerical**

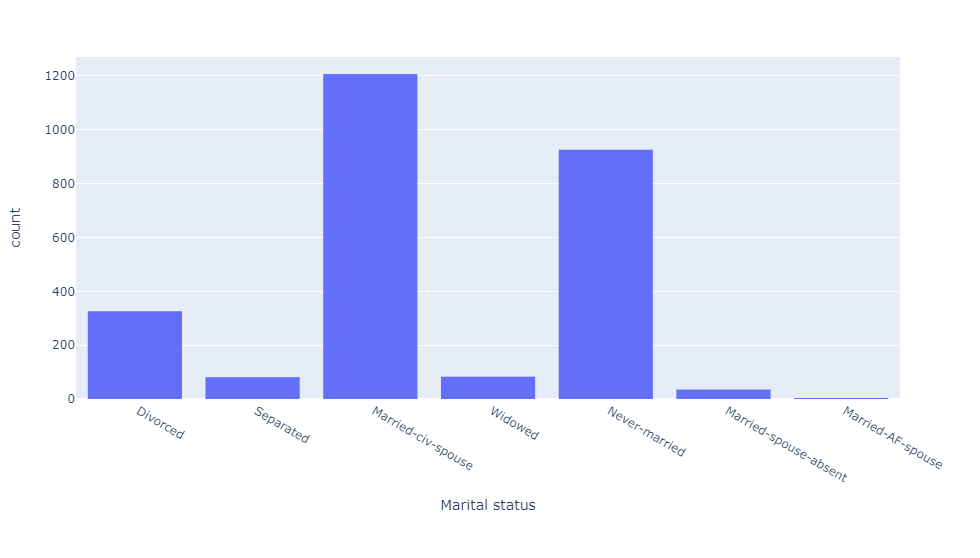
For numerical values I aggregated some statistical data in a table. I also created box plots of all attributes to present the percentile and range statistics.

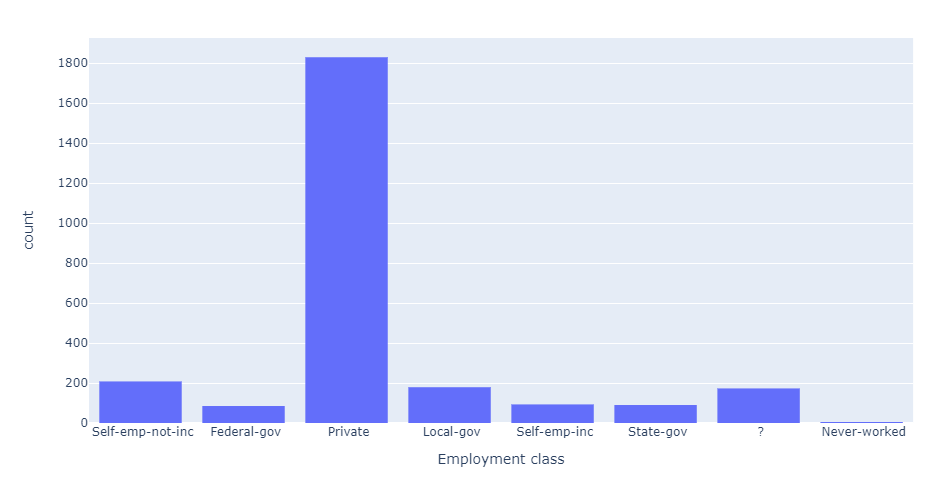


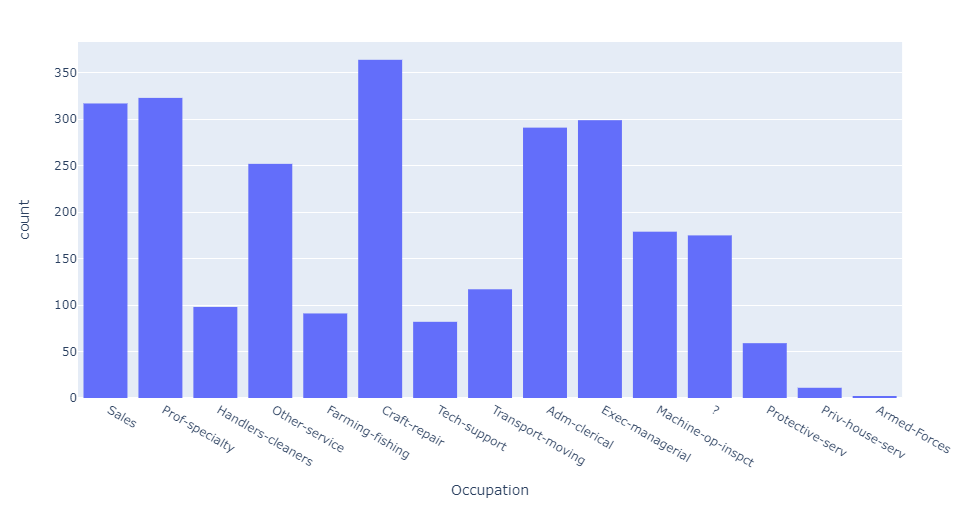
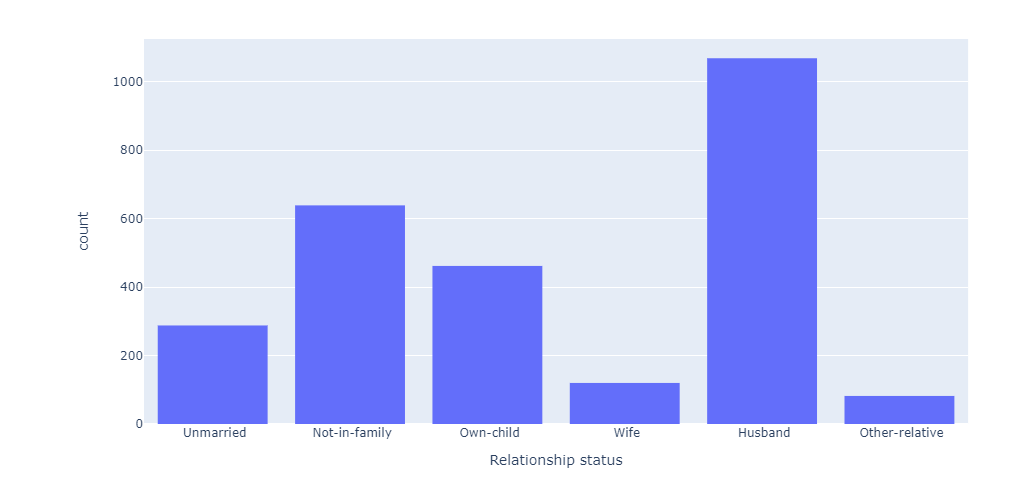
  


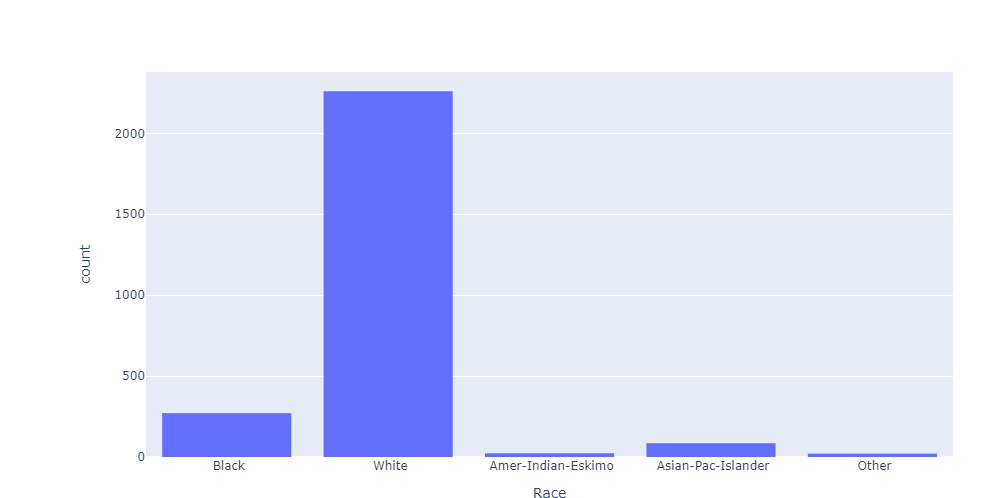
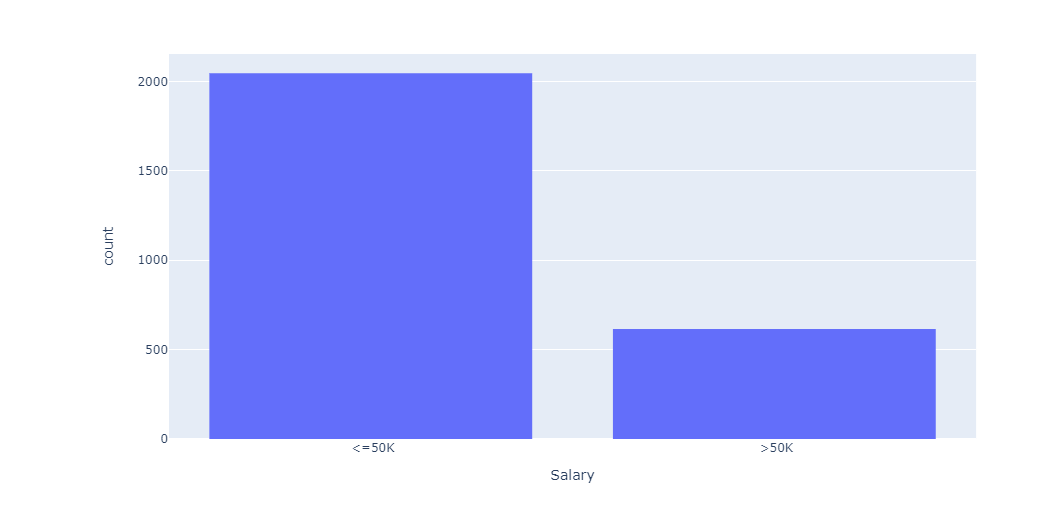
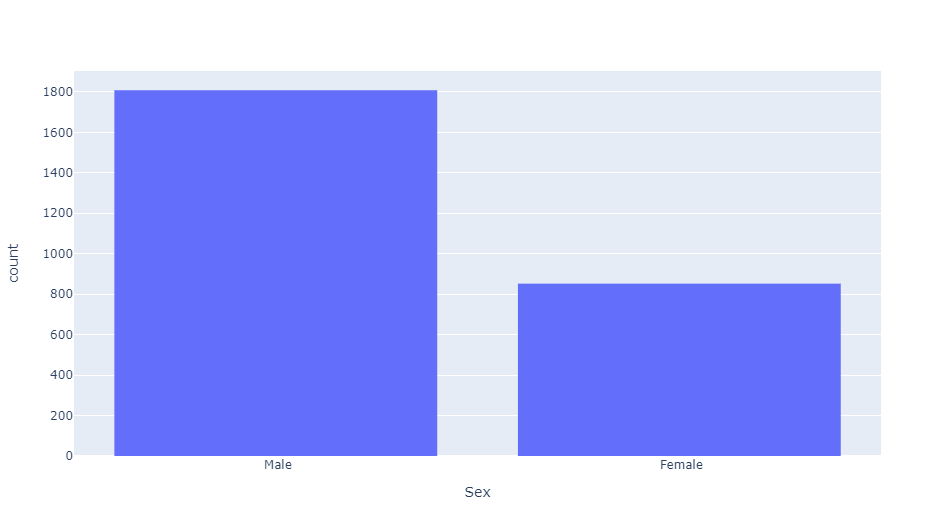
**Nominal**  
For nominal values I aggregated statistics into a table. I also provided a histogram to give an idea of the distribution of values.

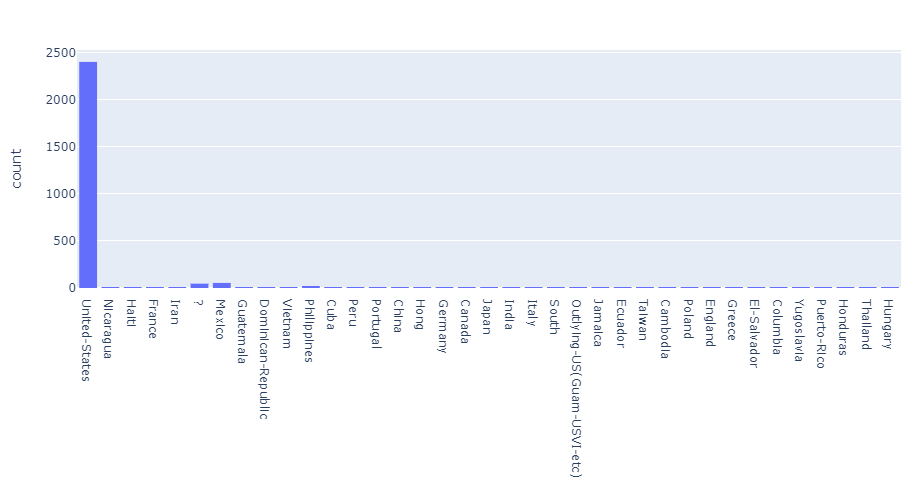










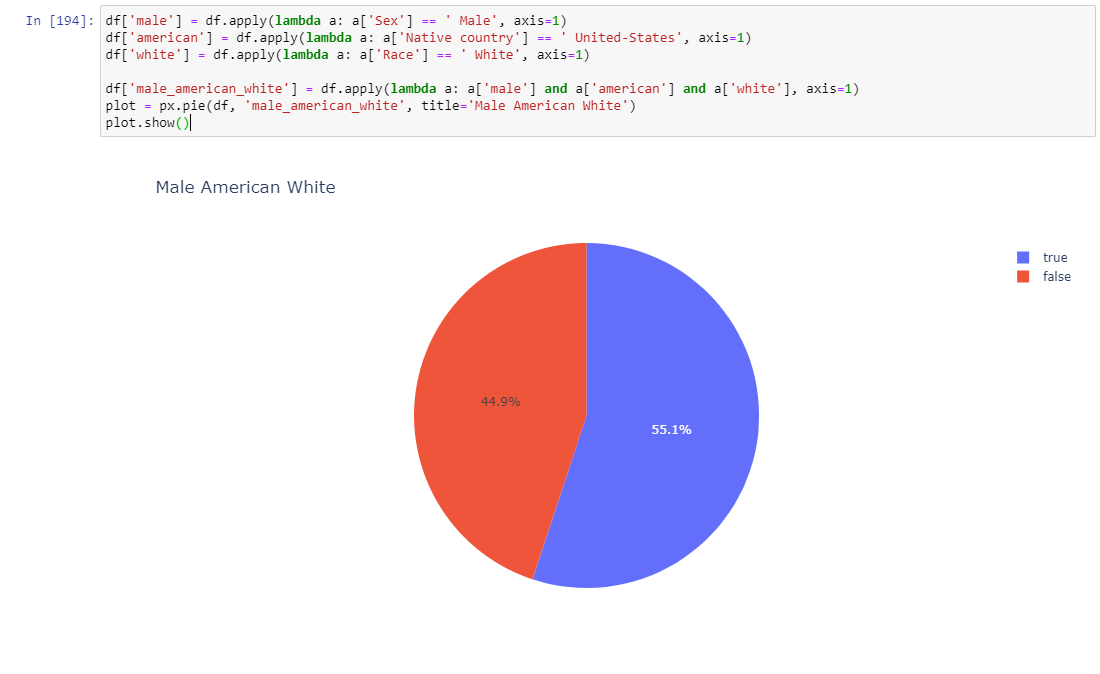


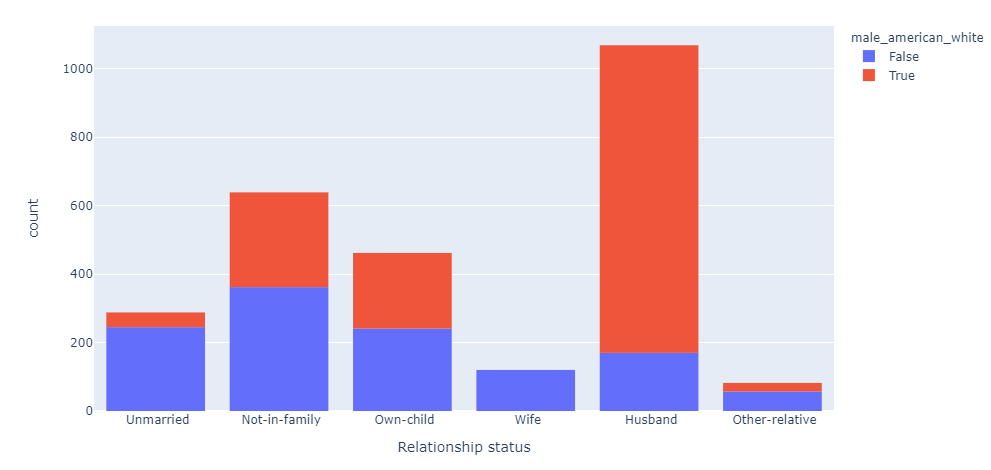
1. **Additional Exploration**

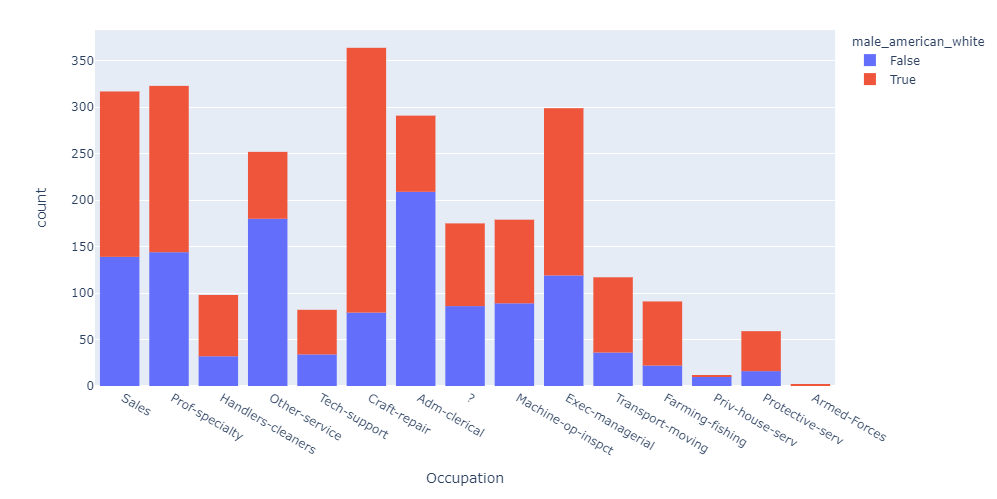
To start with I previously have identified that this data is likely from the census data shown in the following link <https://www.kaggle.com/uciml/adult-census-income>. However, just by looking at the attributes one can assume that this is some sort of census and the amount of entries with a native country of America can lead us to suggest that this was taken in America.

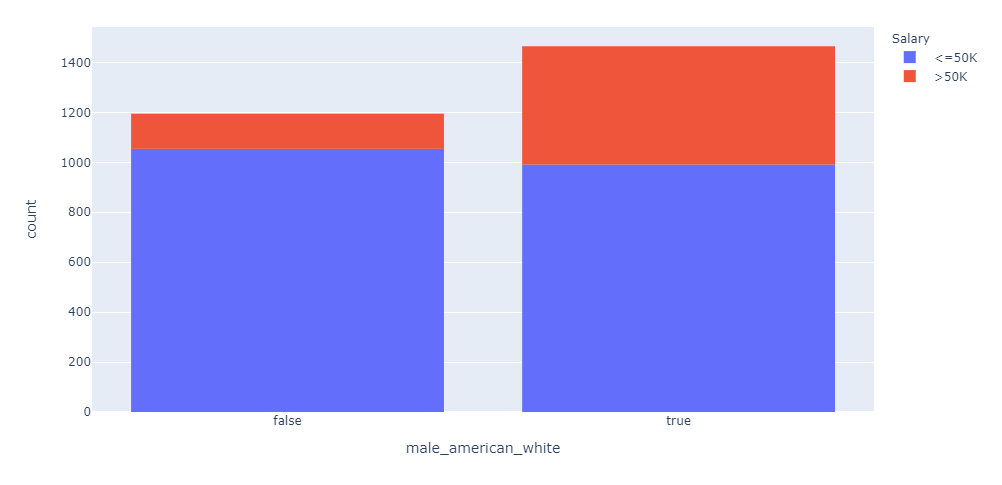
**Male, Americans and White**

Looking at the Mode values for country, race, and gender it is clear to see that majority of the entries fall into these categories. I binarized the combination of these figures to produce the Pi-chart below.



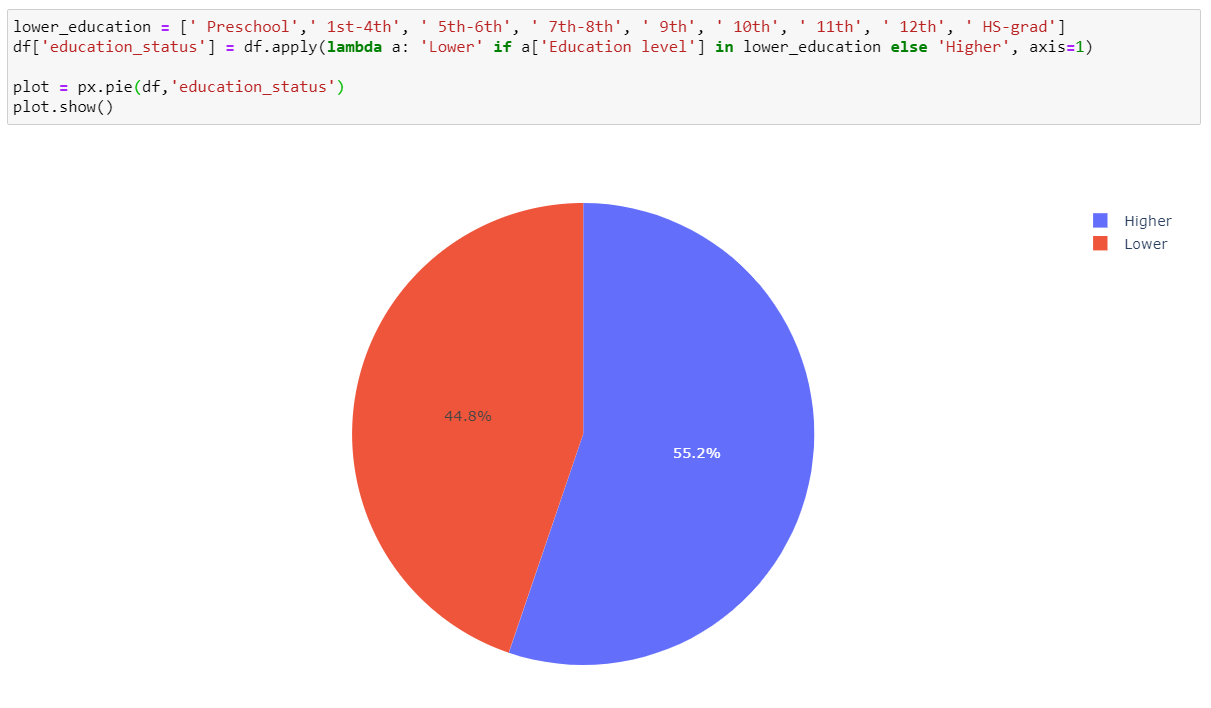
While combining this does not tell us much, it can help gives us some ideas of how this cluster may impact some other attributes. It is safe to say that this category will be the driver for many of the leading nominal attributes. For example, you can come to conclusion that craft-repair is the mode of occupation as it is the most found in the majority group in the study.

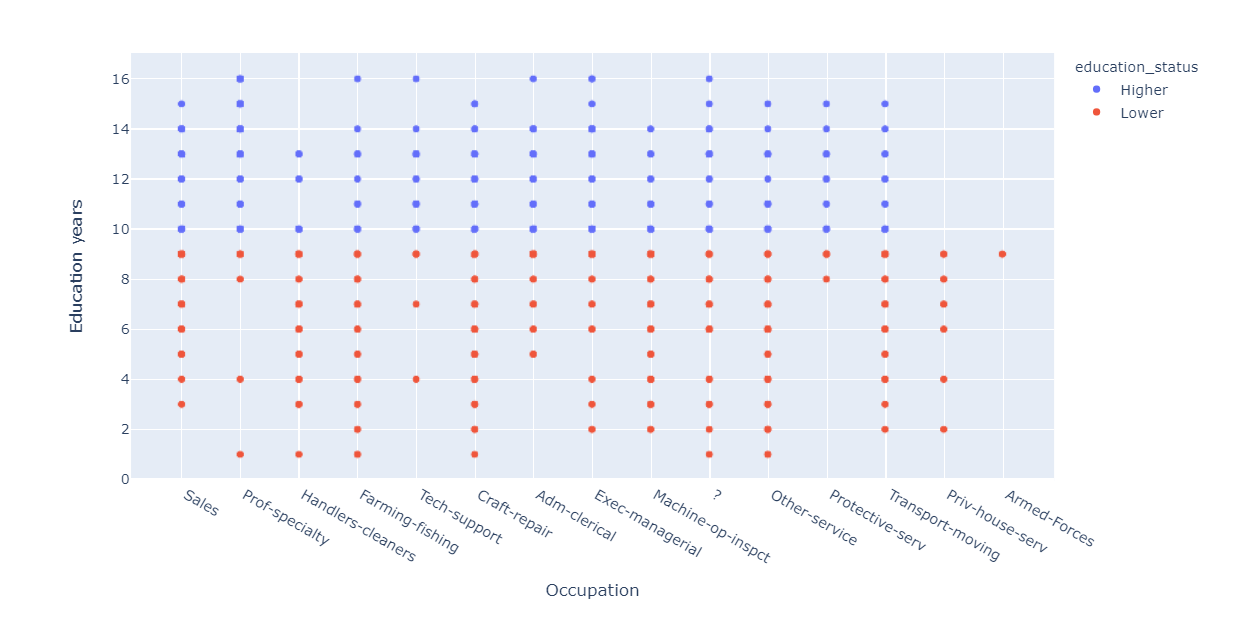


Looking at the salary even though this demographic makes up 55.1% of the total entries they hold 77.07% of the entries that earn over 50k. 

**Education Level**

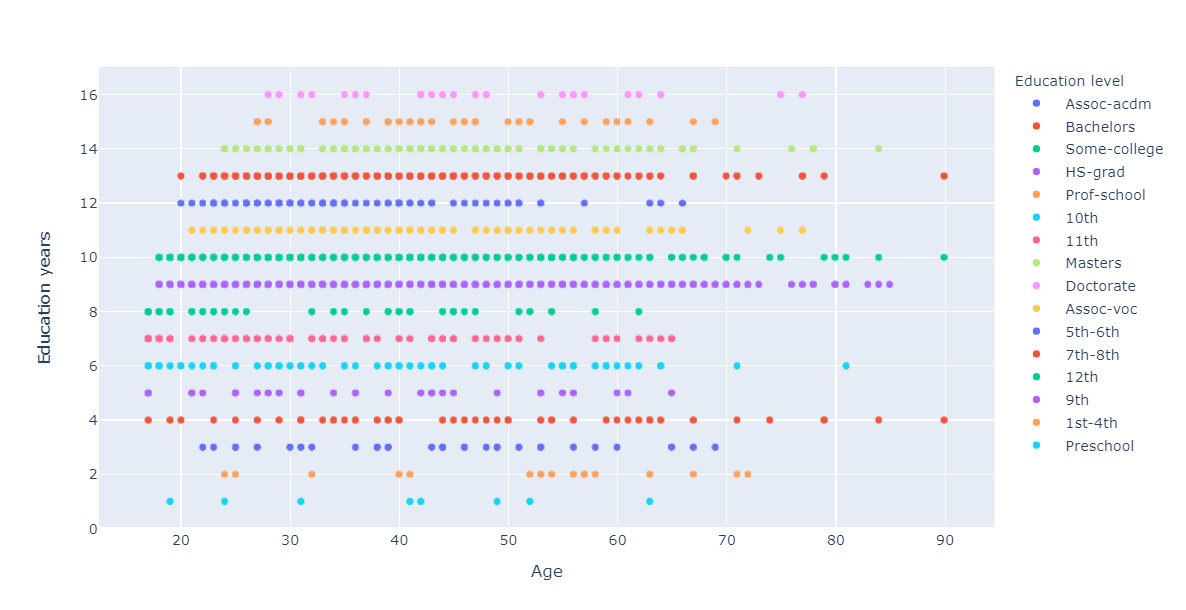
As previously discussed, I would like to break education into groups to turn this from a nominal attribute into an interval. This way we can determine whether there is a correlation between years of education and level and how this effects your job selection, salary, and capital gain/loss. To make this simple I am going to sight every leading up to high school graduate as lower education and everything else as higher. The split between the two is 44.8% in lower and 55.2%



The below graph highlights how the education years evenly splits higher and lower education status, which makes sense as generally the lower levels of education are required to achieve the higher levels. It also highlights how certain occupations are split where protective services are almost entirely higher education and priv-house-serv are entirely lower education. There are some interesting insights where prof-specialty has some outliers in the lower while also lends itself to some of the longest education years of higher. Armed forces also only take lower as I can assume, they require high school graduate who often don’t go on for further education

To investigate some of more information into salary and earnings I have combined the capital gain and capital loss into one column with the following code below. This should show what kind of people make good vs bad investments. Looks as though regardless of education status the capital total clusters around the same spots  



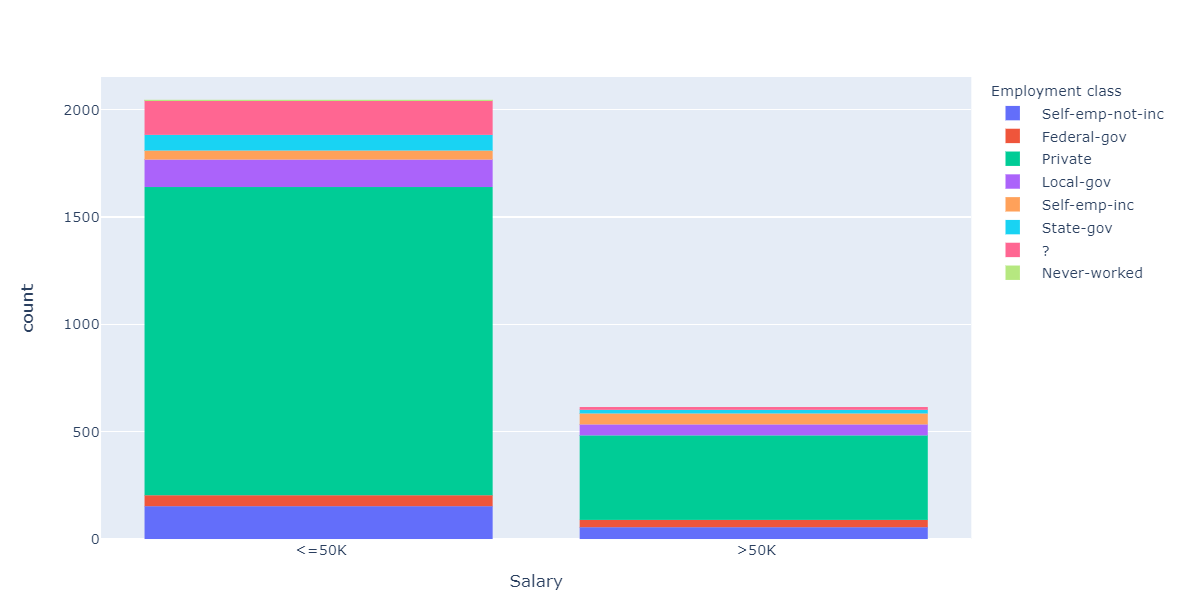

Moving back to the original education level attribute when comparing the average age, education years and education level you can see that it seems to take around 15 years to get a doctorate and generally won’t be reached until your late 20’s.  


**Employment Class**

As previously mentioned, I would like to compare how employment class effects work hours per week. Integrating my previously defined education status also reveals a few interesting clusters. Appears most people work in the private sector and majority within that have a lower education. Self-employed not incorporated has an even spread of work hours have the highest variance of 330.309550. It also is dominated predominantly by lower education entries.



Looking at salary against employment class there doesn’t seem to be a clear-cut indication a particular class is more likely be in a certain bracket over another.



**Non-Americans**

In this section I wanted to explore how some of the non-Americans faired in certain values in the data set. Non-Americans make up 9.62% of the entries in my data. In the diagram below you can see that majority of the self-employed are American born. This goes the same for government jobs which are dominated by Americans. It seems non-Americans are clustered in the private sector and working average hours per week.

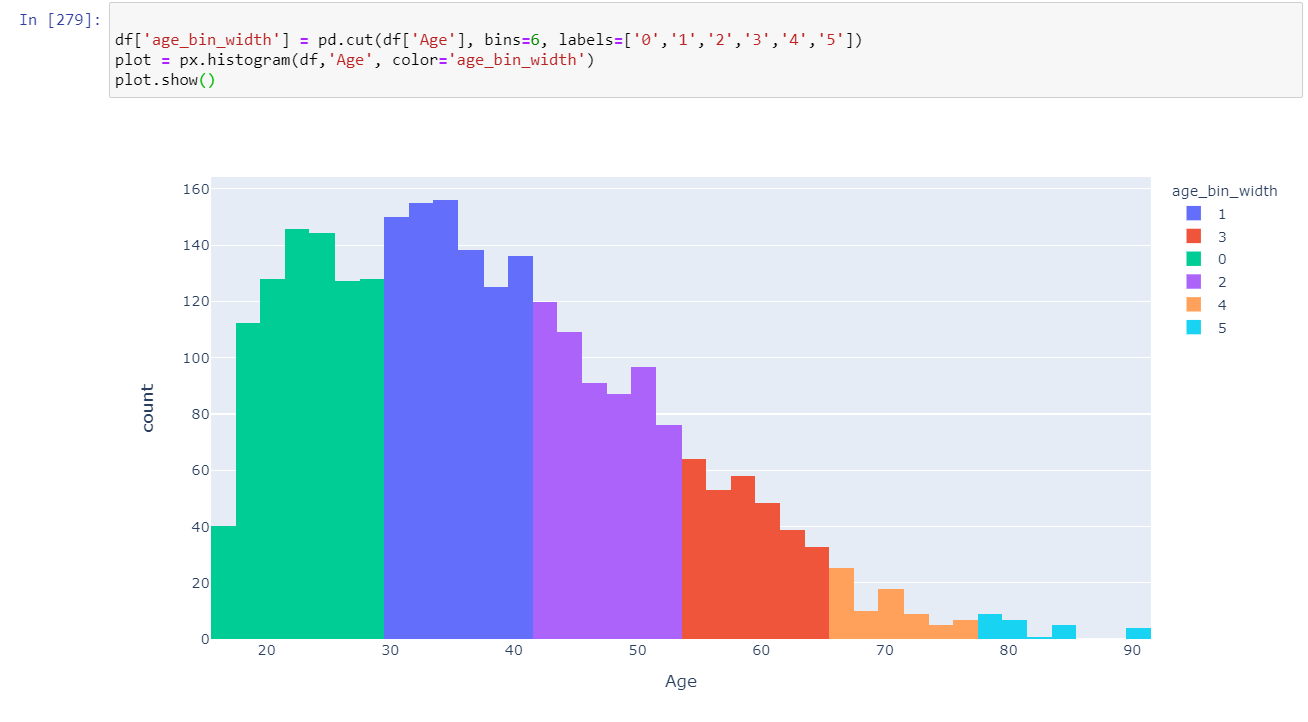


Looking at the salary you can that non-Americans make up a very small portion of the greater that 50K salary. Making up only 5.63% of the total amount in that range.

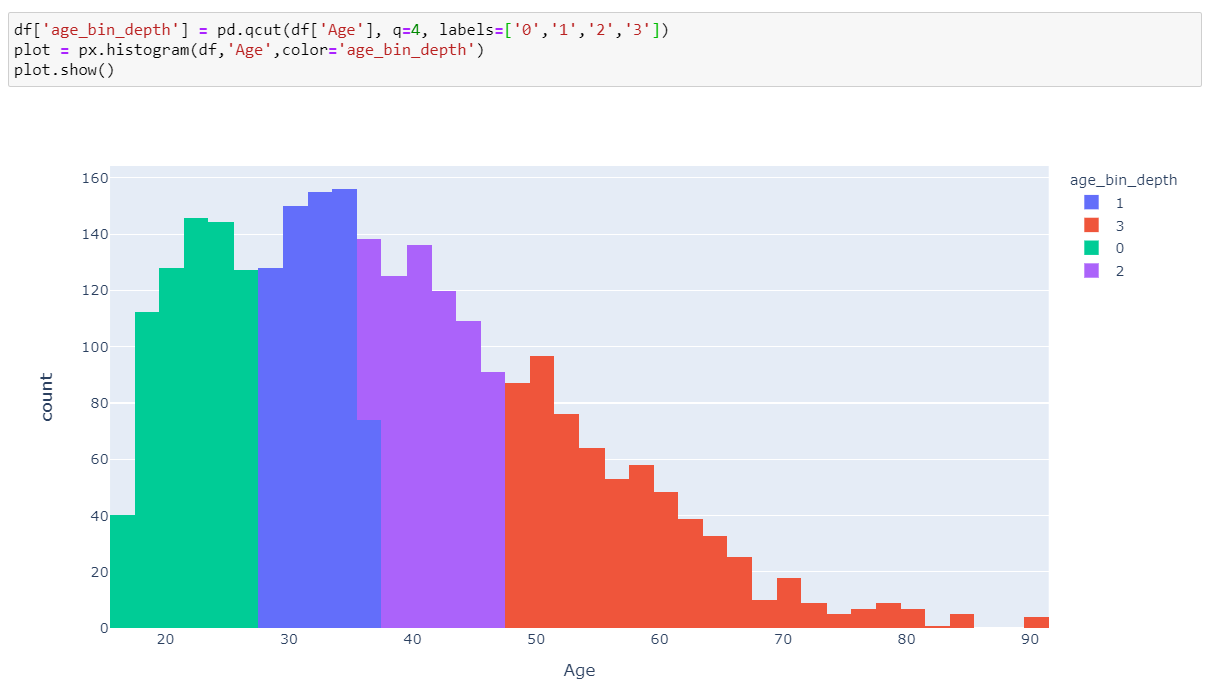
**1B: Data Preprocessing**

1. Binning

Both bins where made with the same general idea in mind. I believe this has a nice split between what could be considered your early, middle, and late stages of life. For equal width, the last two bins have relatively low counts but represent a period of retirement that can be excluded or included for analytics. The end of the first bin represents a period where people can reach a master’s level of education. This means you can consider group 0 as a developing period for education level compared to group 2. Equal depth follows this to an extent except the last few bins have been combined into one which would be useful if you wanted to included outliers of the age range in calculations.

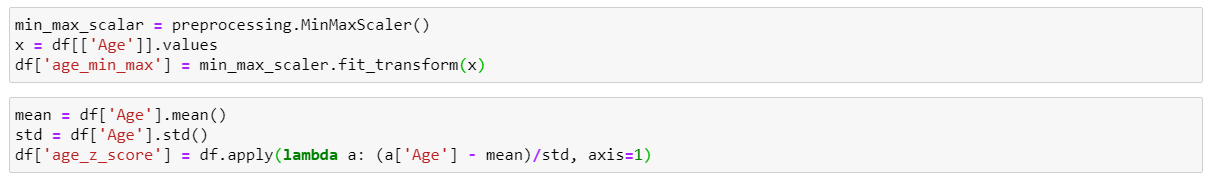


**Equal Depth**



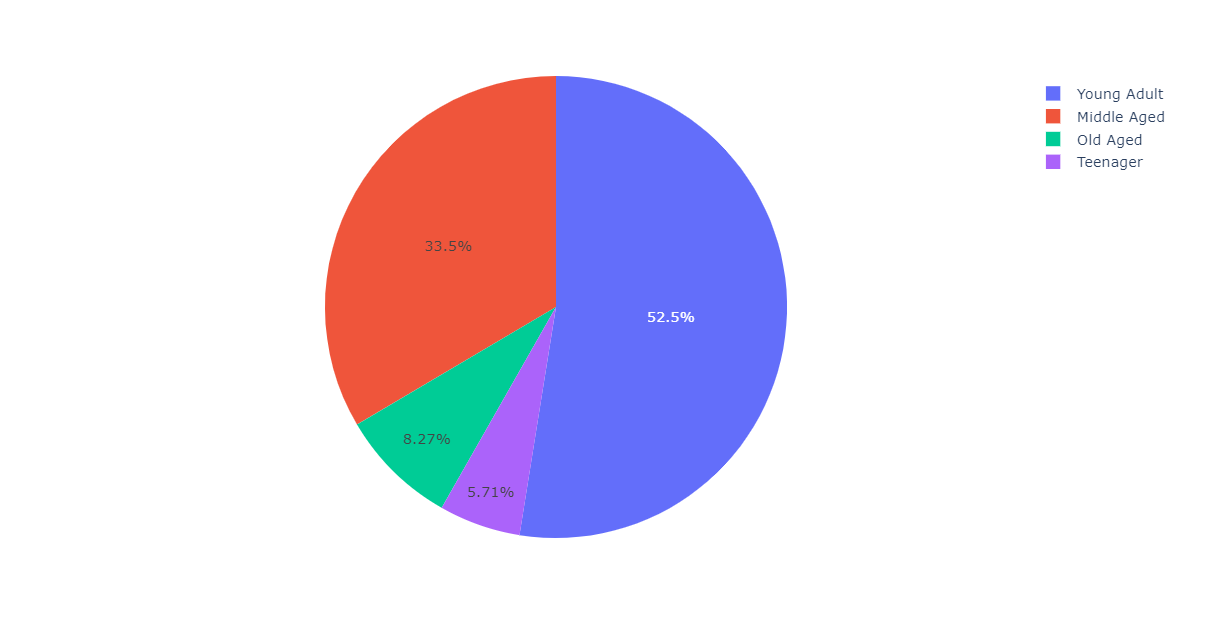
**B. Normalize**

Both methods of normalization were done in python as shown below. The min max scalar utilized the ‘sklearn’ library and moved the values with the range of 0->1. The age z score values where defined using the z score formula of (value - mean) / (standard deviation).



**C. Discrete age**

To create the discrete age groups, I defined a function which would return the value of the group given a certain age. I applied this to a new column of the data frame and showed the frequency in a pie chart shown below. Note: there are no children in my data.



**D. Binarize Married Status**

Binarizing the marital status is fairly easy using the Pandas library as it has a get\_dummies function to help break the values down into binary figures. This function does as expect and converts categorical data into binary data.



**1C. Summary**To conclude here are a summary of the insights I have found in the data. My main focus was looking at how education level effected the other attributes.

Majority of the entries fall into a Male, White and American category taking up 55.1% of the total. In terms of salary over 50K this classification holds 77% of the entries. Additional some entries such as relationship status and occupation are heavily linked to the dominant nature of this category.  
  
Splitting education into lower (high school graduation and lower) and higher splits the entries evenly. This leads to also identifying that higher education requires going through the years of lower education. Protective services, specialty professions and tech support are held mostly by the higher educated while private house services and armed forces are held by the lower educated. It also appears that majority of education development occurs in people 20’s with the earliest that someone can attain a doctorate appearing at the age of 28.

The self employed is largely taken up by the lower educated who have a large variance in the number of hours they work per week. Most people work in the private sector with government jobs being reserved for the higher educated. Non-Americans also work mostly in the private center and while holding 10% of the total entries only hold 5% of the salaries over 50K.

Moving forward I think there is a high correlation into how education effects the populations Occupation and salary. Further investigate how certain demographics have access to the requires years of education to reach the higher levels could create some interesting findings.