CSCI4150U: Data Mining
Lab 02

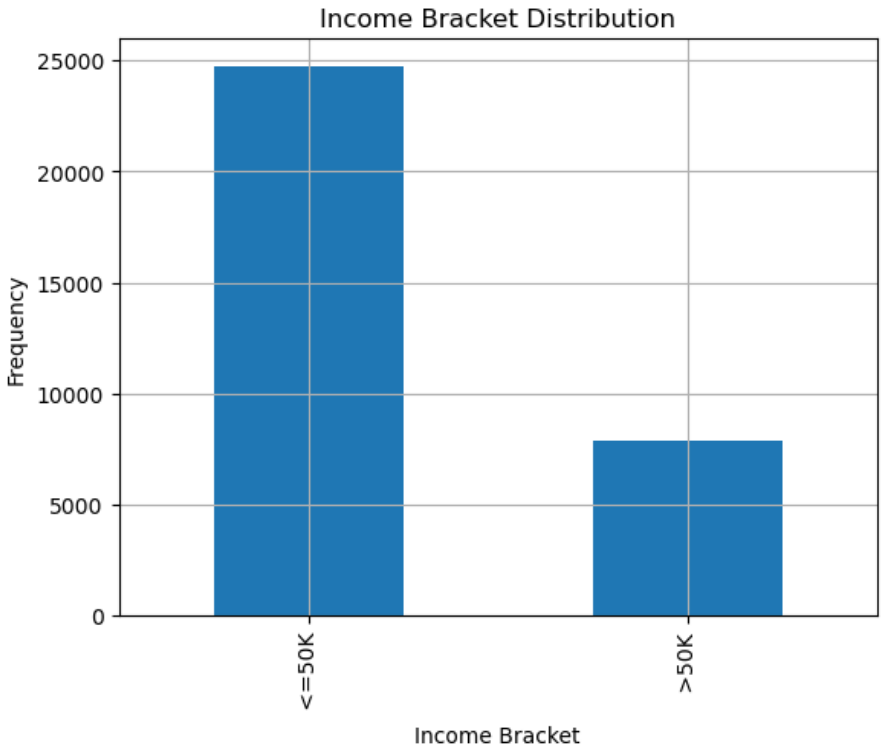
Syed Naqvi
100590852

September 29, 2024

Part I:

3.

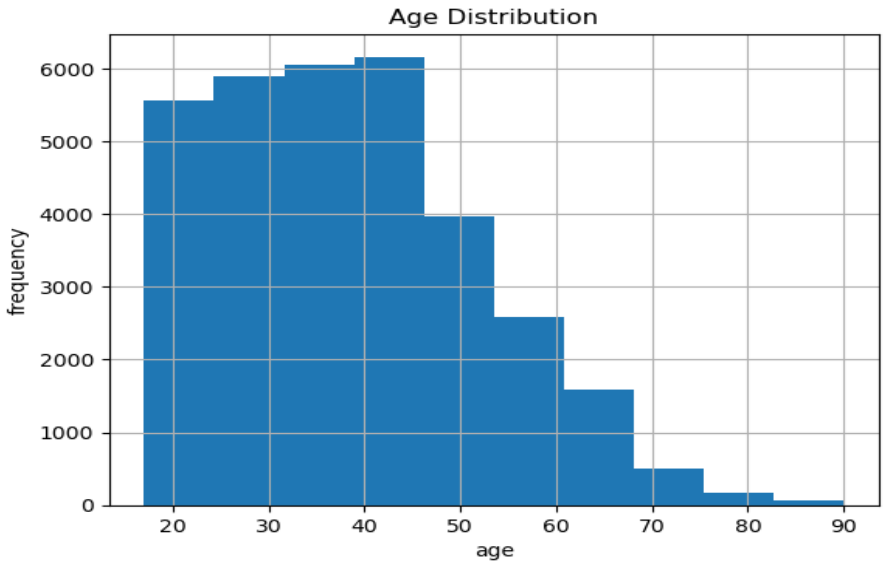
Since the class variable is a discrete feature, we can observe its distribution using a bar plot:



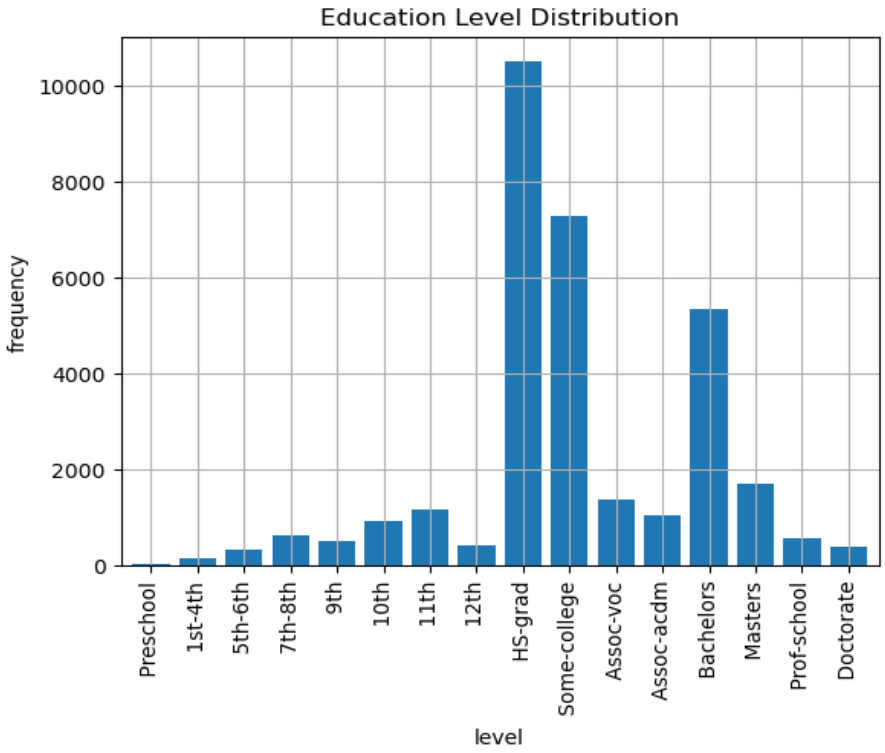
The diagram clearly shows that the majority of surveyed people (approximately 250000) earn less than or equal to 50k per year. This is roughly 3 times as many people (approximately 75000) who earn more than 50k per year.

4.

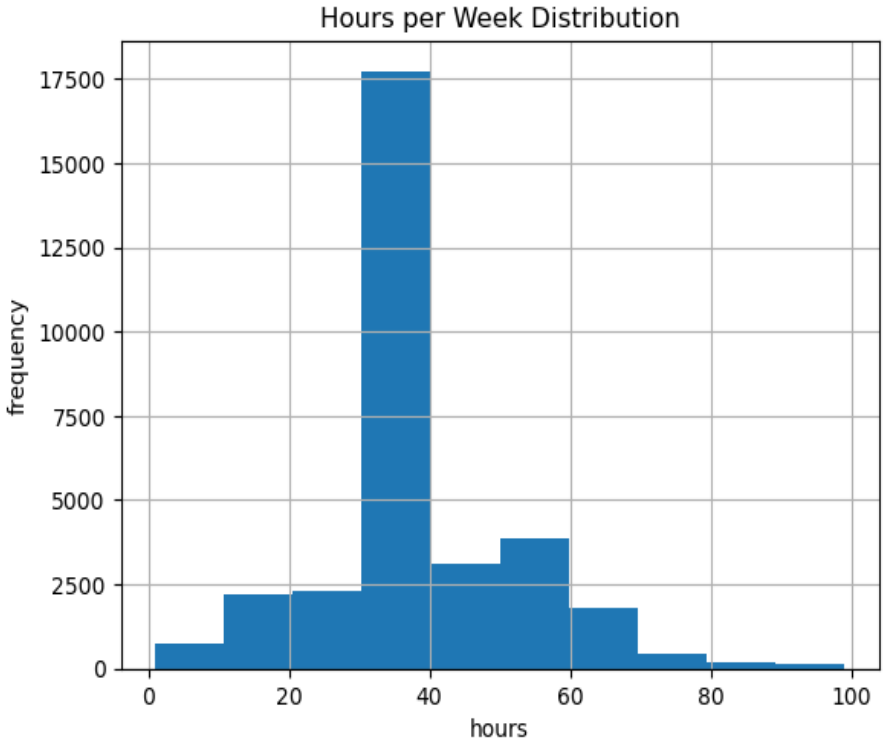
Let us visualize the distributions of population age, education level and hours worked per week:



From the age distribution we can see that the population mainly consists of more youthful individuals, with the overwhelming majority of people less than 50 years old. A younger demographic likely explains much of the over-representation of lower income individuals since income usually increases with age.



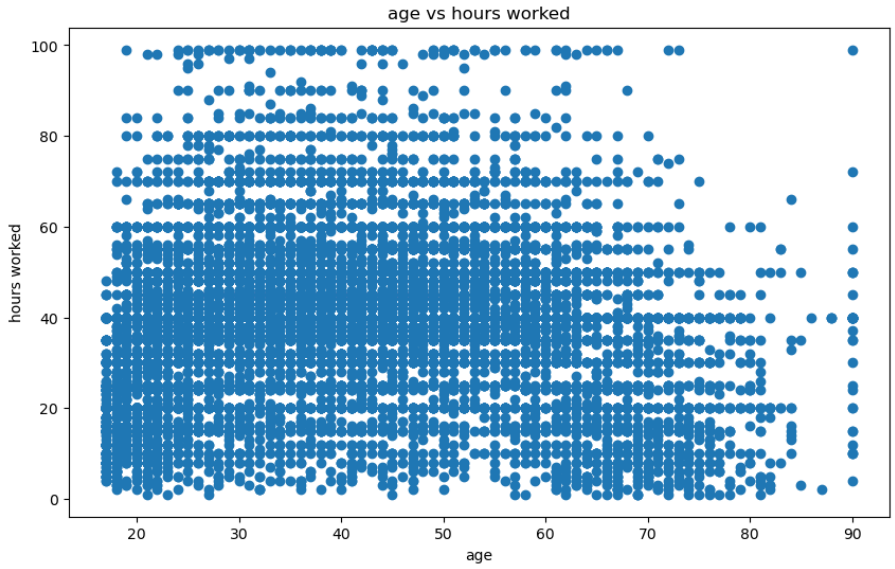
The most common levels of educational attainment seem to be high school graduates, then some amount of college followed by bachelors degrees. The majority of people having less than a bachelors degree of education while also being relatively young further explains the higher degree of low income earners.



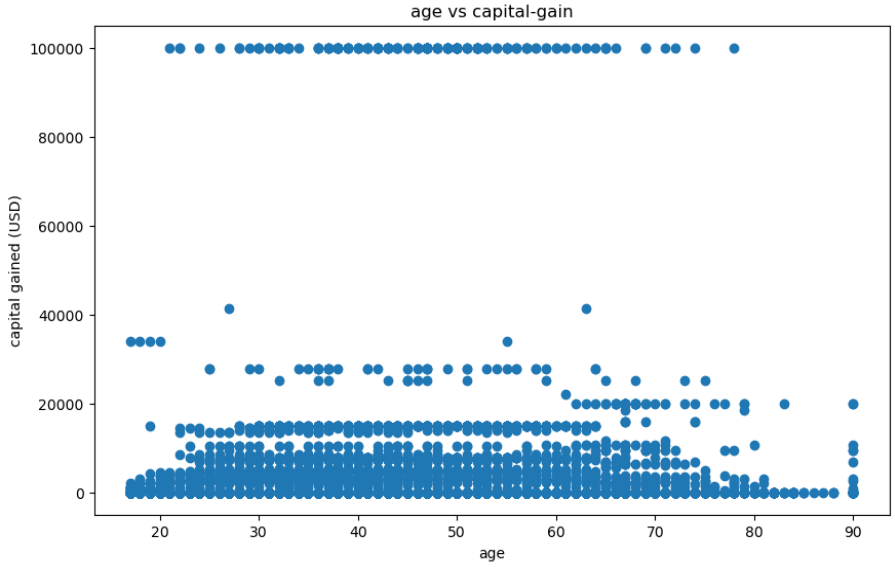
The overwhelming majority of people seem to be working less than or equal to 40 hours per week. This indicates a higher likelihood of 9-5 employees in the surveyed population as opposed to self-employed people or entrepreneurs who often work longer hours with higher earnings.

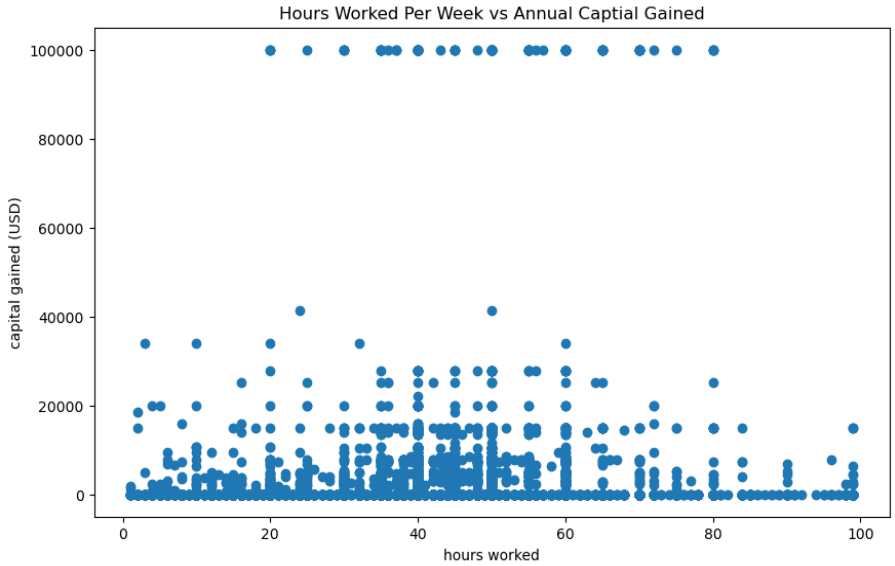
5.

The following plot explores potential relationships between age and hours worked:



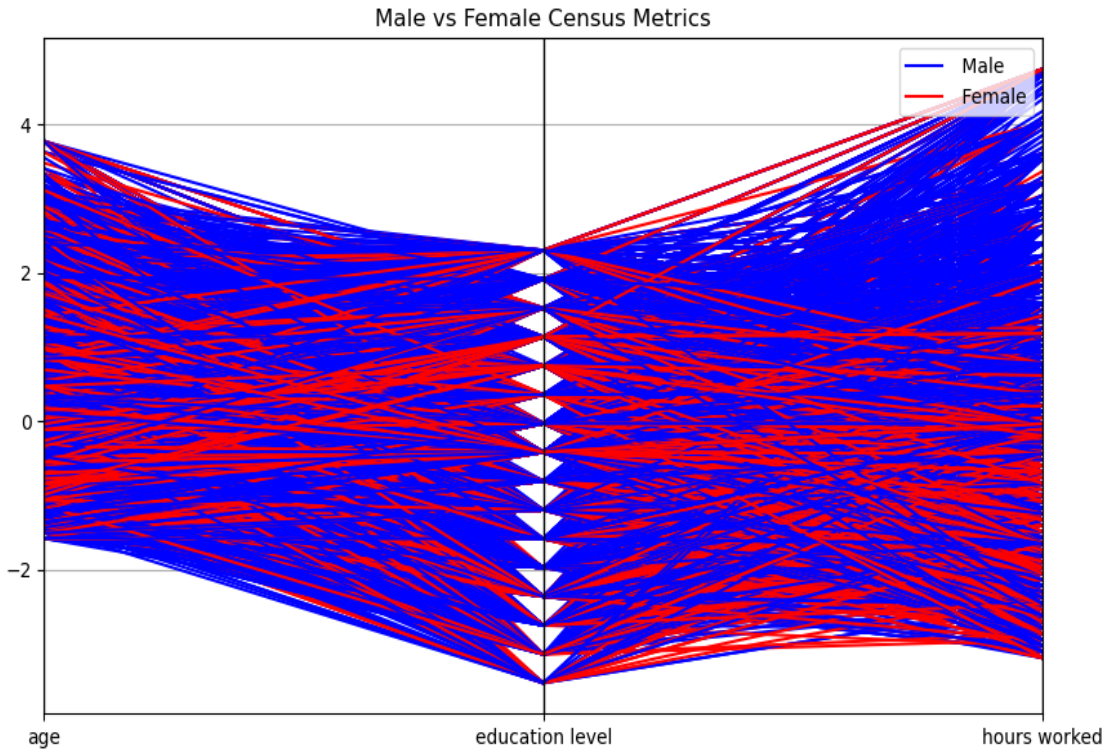
While I had expected to see a stronger negative correlation where people work fewer hours as they age, there does seem to be some alighnemt with that notion in the way of a parabolic relationship. We can see that the majority of people in their late teens and early twenties start working with great variability in their number of hours. I imagine this is likely due to a mixture of part-time students and full-time entry level workers. The variability in the number of hours certainly appears to stabilize around 45 hours per week throughout middle age from late twenties to about 60. At this point people likely begin to retire or greatly reduce their workloads resulting in a more downward trend.





Exploring the relationship between age vs captial gain as well as hours worked vs captial gain, we see very simlar distributions. It appears that the captial gains remains relatively stable throughout people's adult lives with a slight parabolic curve suggesting the majority of gains come around the 3rd quarter of life. It may be the case that with age come more accumulated assets that can be sold as well as better business sense. The hours worked vs captial gains plot appears very similar in shape where the majority of captial gains are made be people working roughly 50 hours per week. Perhaps working fewer hours than this means people are not generally wealthy enough to afford assets to sell and beyond this level of work there is not much time to invest in other affairs besides work. There does seem to be a set of outliers in the plots involving capital gains where a small subset earns nearly 100000 USD per year.

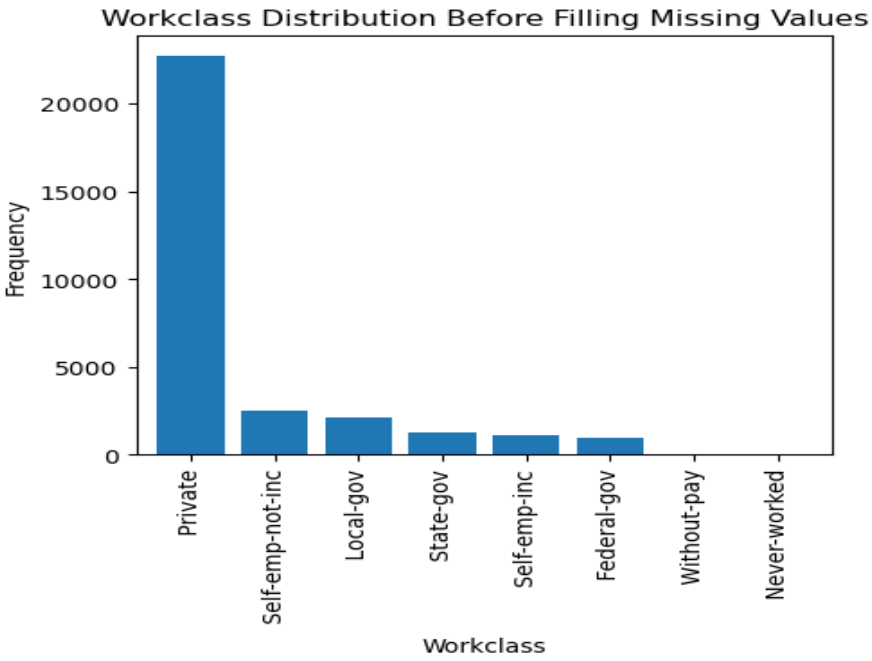
6.

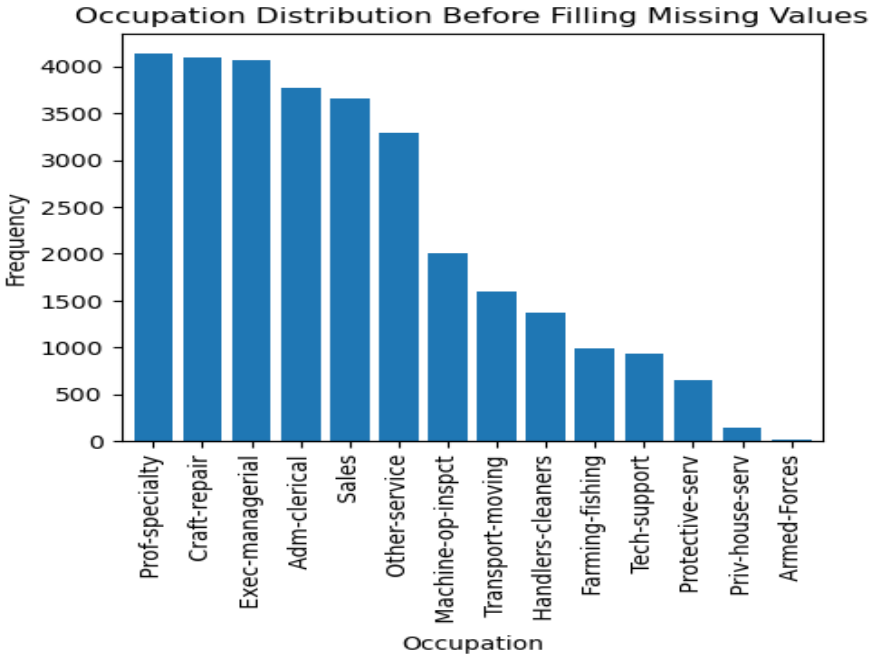


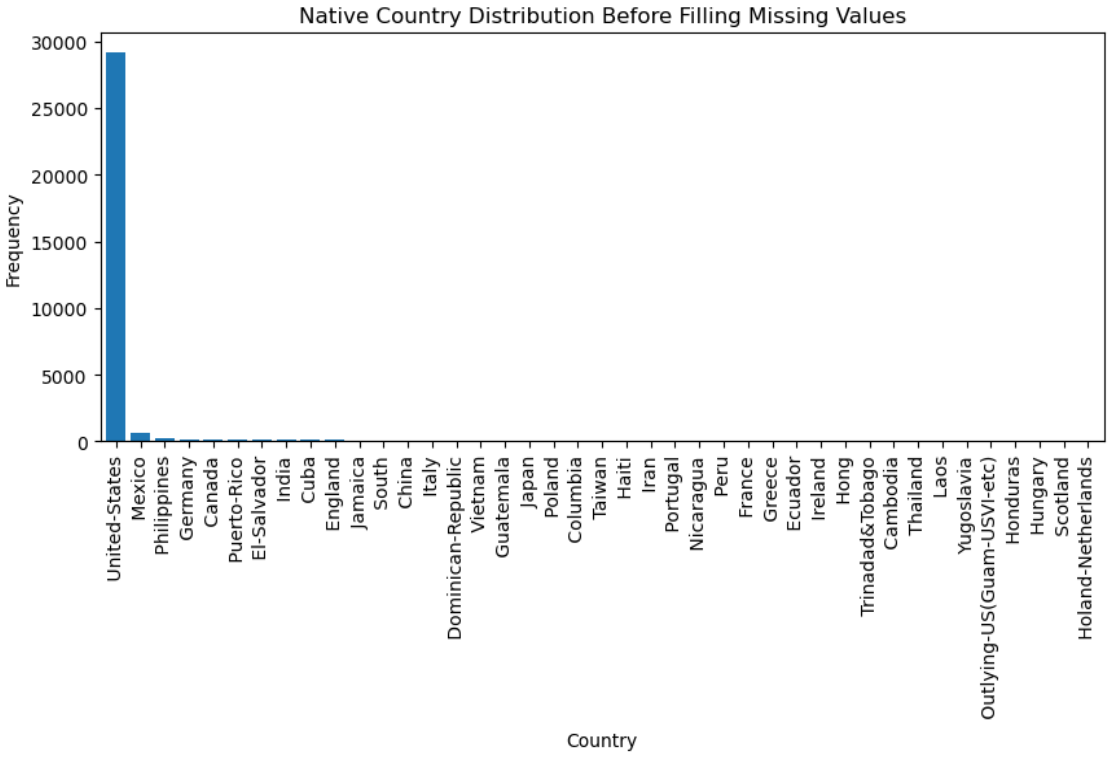
The above figure contains standardized age, education level and hours worked axes in order to make the parallel graph more readable and patterns more apparent. We can see that males and females are genearlly represented equally across all metrics other than hours worked where we can clearly see that males are over-represented among the people with the most hours worked. There also seems to be a greater proportion of men among the upper and lower levels of education while women occupy more of the middle levels. I suspect the number of hours worked likely contributes to a subsequent gender pay gap.

Part II:

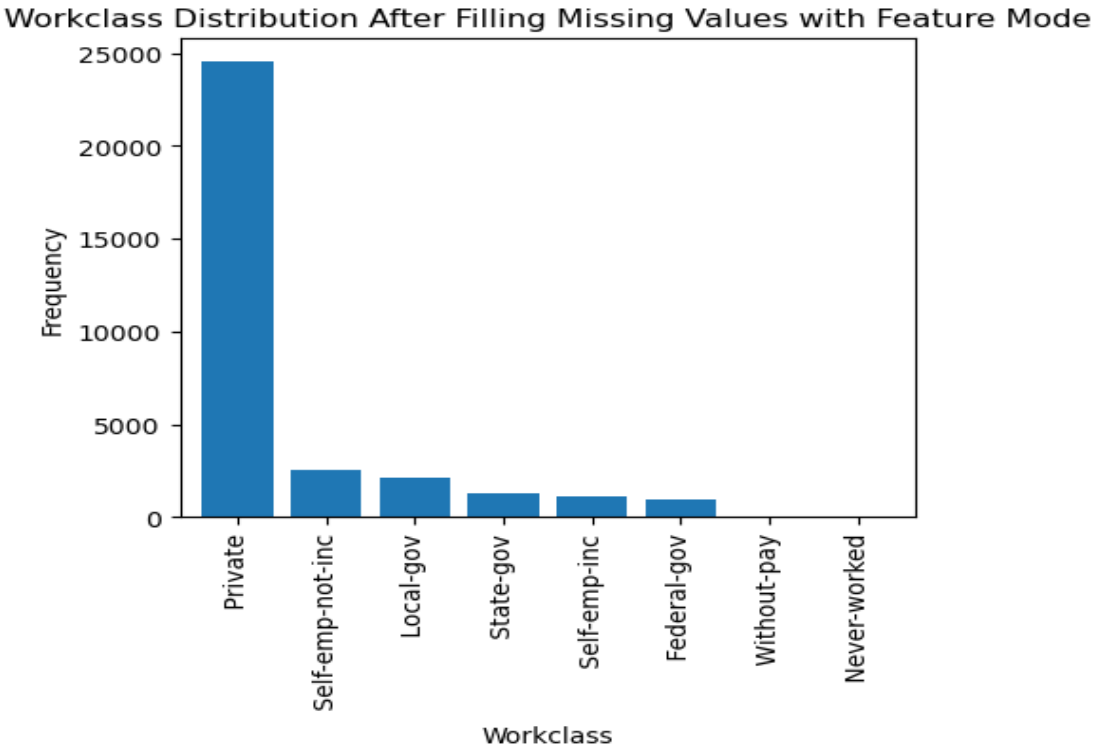
3.

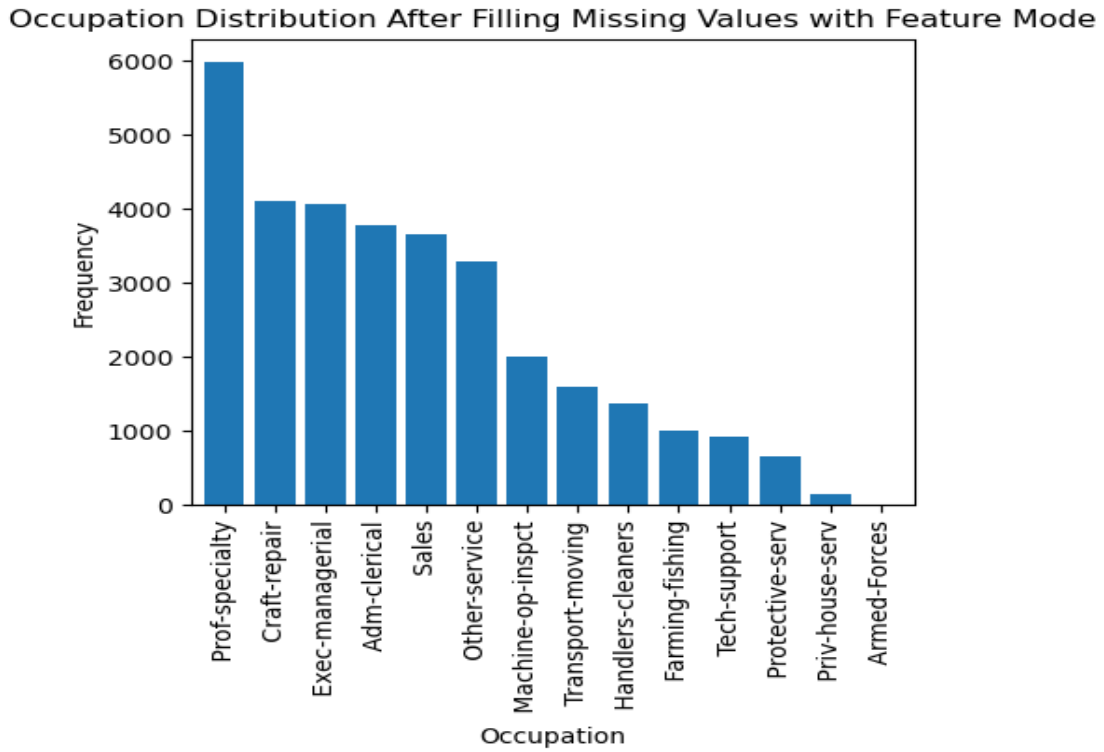

Workclass Distribution Before Filling Missing Values
Frequency
20000
15000
10000
5000
0
Private
Self-emp-not-inc
Local-gov
State-gov
Self-emp-inc
Federal-gov
Without-pay
Never-worked
Workclass

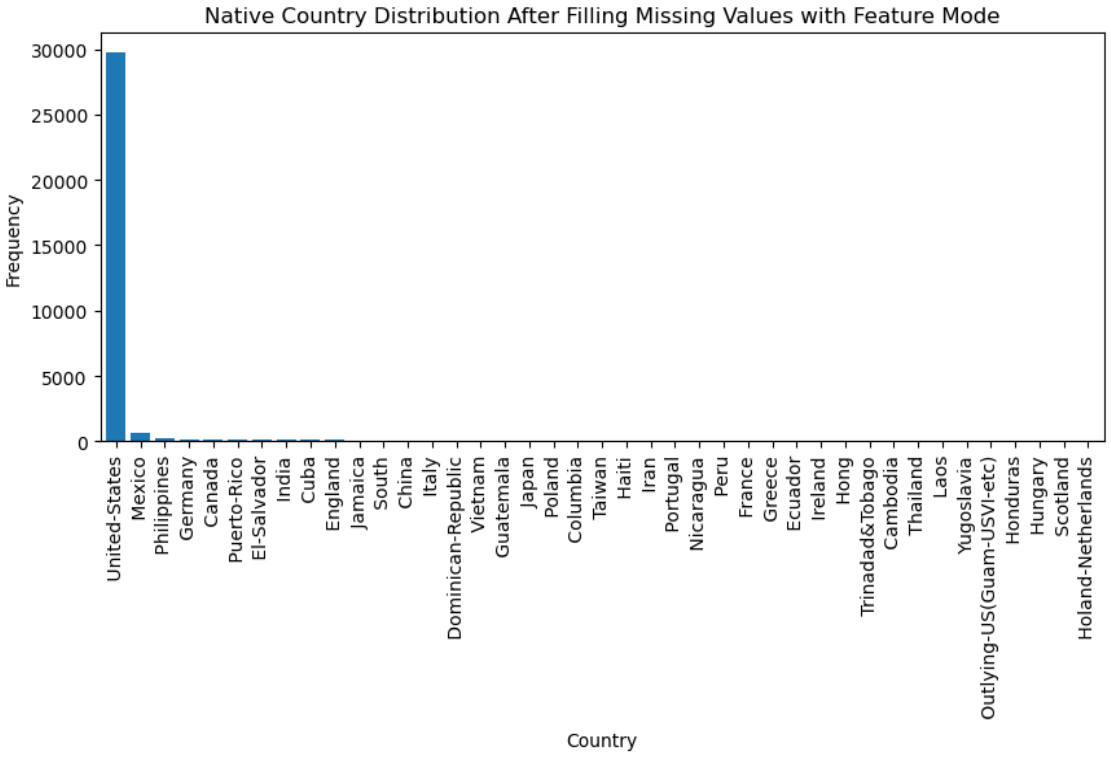

Occupation Distribution Before Filling Missing Values
Frequency
4000
3500
3000
2500
2000
1500
1000
500
0
Prof-specialty
Craft-repair
Exec-managerial
Adm-clerical
Sales
Other-service
Machine-op-inspct
Transport-moving
Handlers-cleaners
Farming-fishing
Tech-support
Protective-serv
Priv-house-serv
Armed-Forces
Occupation



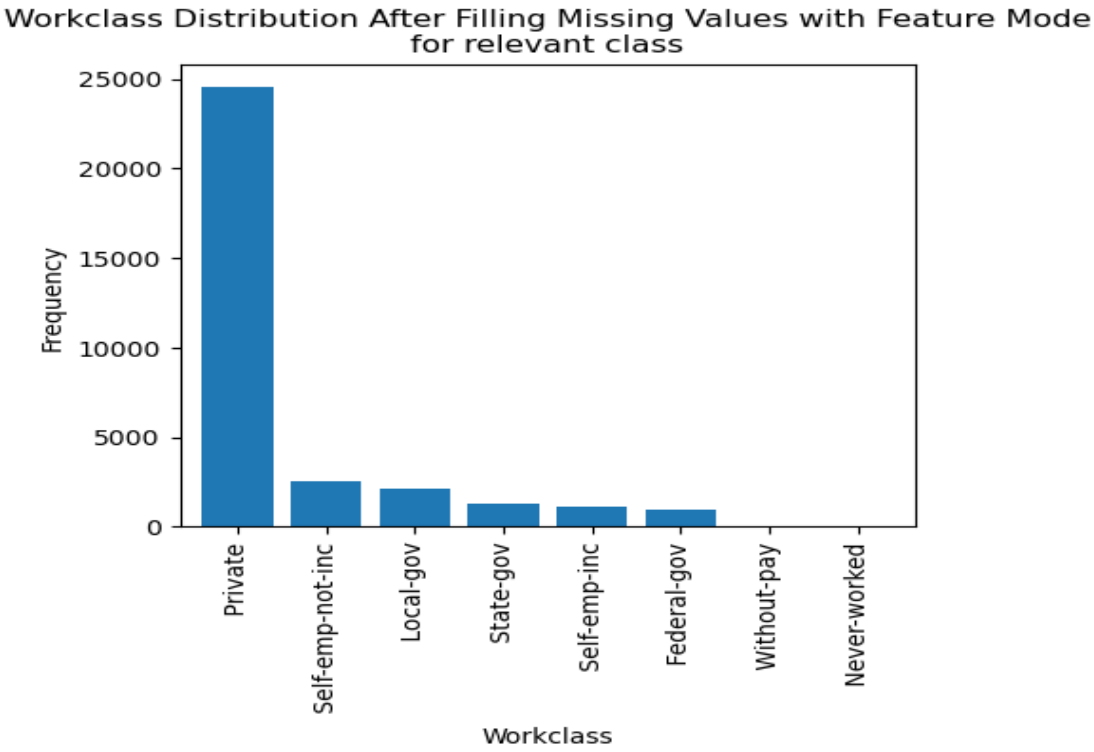
The above three graphics visualize the distributions of the three features with NaN entires. Each feature was discrete and so its distribution was visualized using a bar gprah. For the above plots, records with missing entries were simply ignored.

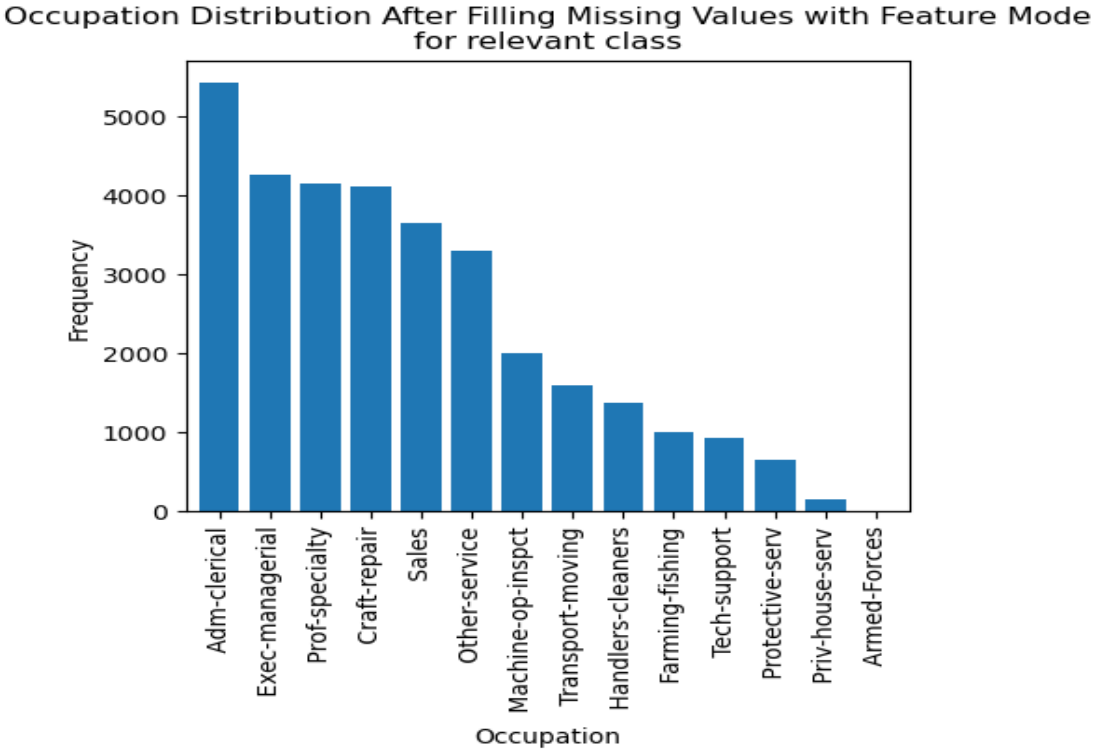


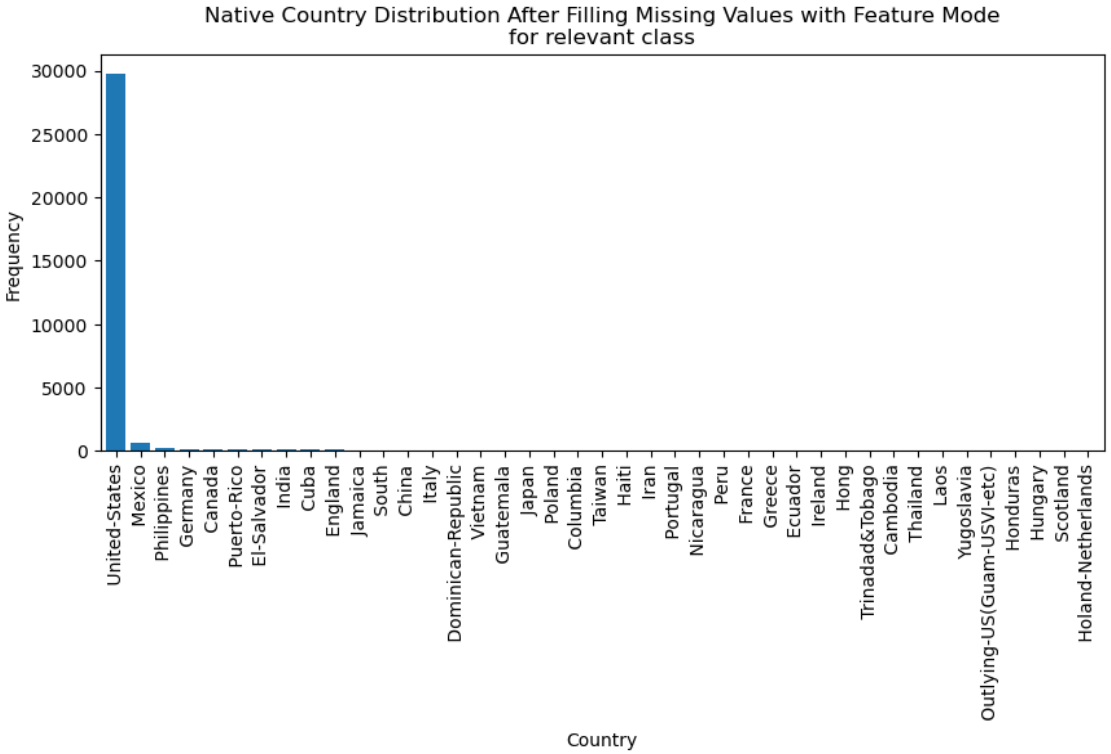




The above three graphics were generated after replacing NaN entries with the mode of their column. The effects of this are clearly reflected in the diagrams where the most frequent value in each distribution experienced a significant increase since all missing entries were set equal to this value. In the case of prof-specialty occupation, this value jumped from near 4000 to near 6000.


Workclass Distribution After Filling Missing Values with Feature Mode for relevant class
Frequency
25000
20000
15000
10000
5000
0
Private
Self-emp-not-inc
Local-gov
State-gov
Self-emp-inc
Federal-gov
Without-pay
Never-worked
Workclass


Occupation Distribution After Filling Missing Values with Feature Mode for relevant class
Frequency
5000
4000
3000
2000
1000
0
Adm-clerical
Exec-managerial
Prof-specialty
Craft-repair
Sales
Other-service
Machine-op-inspct
Transport-moving
Handlers-cleaners
Farming-fishing
Tech-support
Protective-serv
Priv-house-serv
Armed-Forces
Occupation



We now see a slight decrease in some of the most frequent values since they are no longer necessarily the most frequent for their feature depending on the class of the instance. The missing values are thus being distributed among the top most frequent values, although the change is not too significant.