Basic Data Visulization and Main Insights

Arnold Jiadong Yu

March 7, 2019

Before visulizing the relationship between each attribute and the class label. I have read adult.name.txt and old.adult.name.txt, and gained a basic understanding of the dataset.

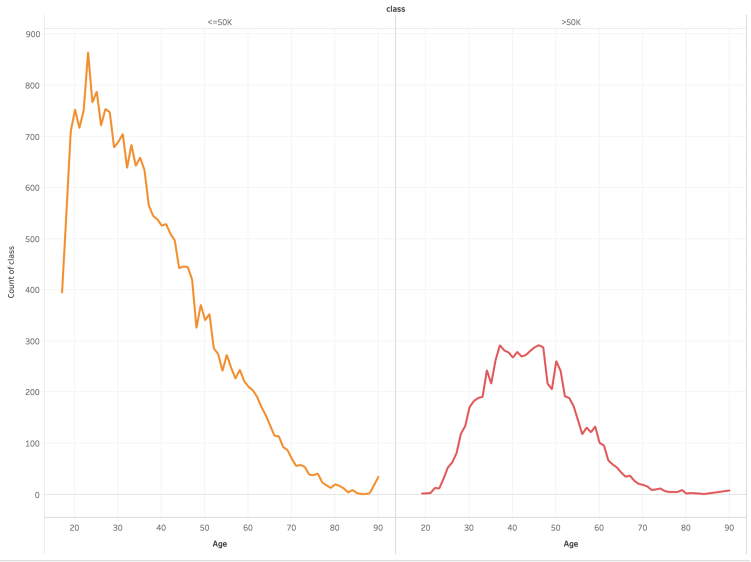
First, I import the dataset into Tableau and observed the following:

1. There are 6 continuous and 9 categorical attributes.
2. All missing values are in the attributes of work-class, occupation, and native-country.
3. Education and Education-num means the same thing in the dataset. So remove education-num attribute.
4. The number of data in $\leq 50k$ is 24720 and in $> 50k$ is 7841. By information obtained from social security website[1] , median income was 16118.02 and mean income was 22786.73 in 1994. This showed the distribution is very positive skewed which proved that we should have much more population in the category of $\leq 50k$ than in the category of $> 50k$. The data is a good representation of population. But comes to k-fold, we need a even dataset.

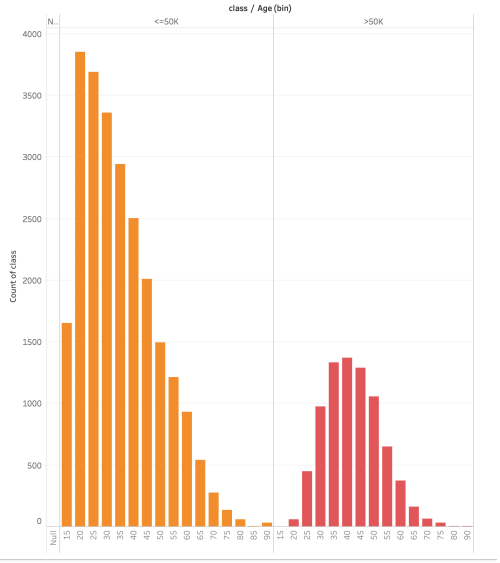
[1] **https://www.ssa.gov/oact/cola/central.html**.

Now, let's look at the relationship between each attribute and the class label.

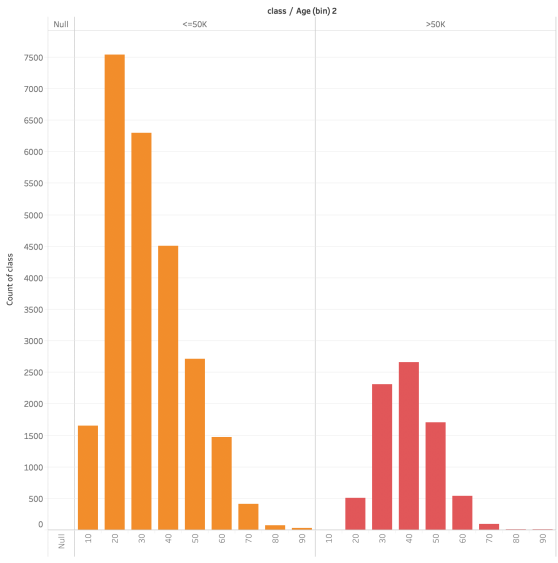
Age



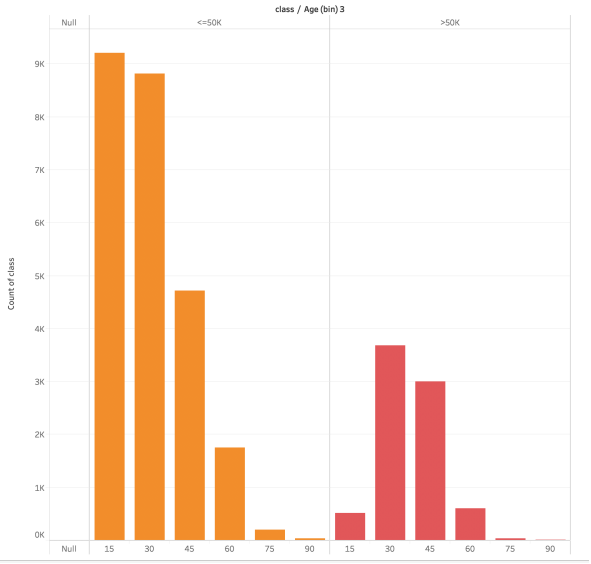
5



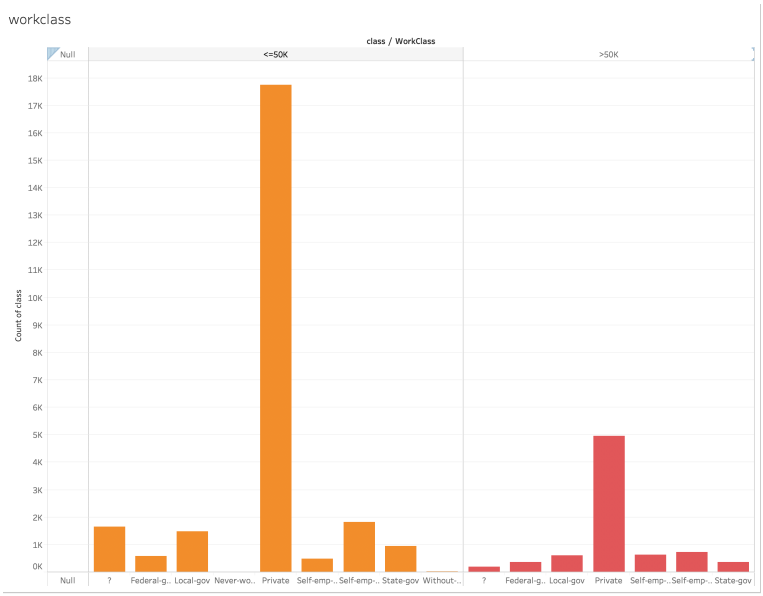
10



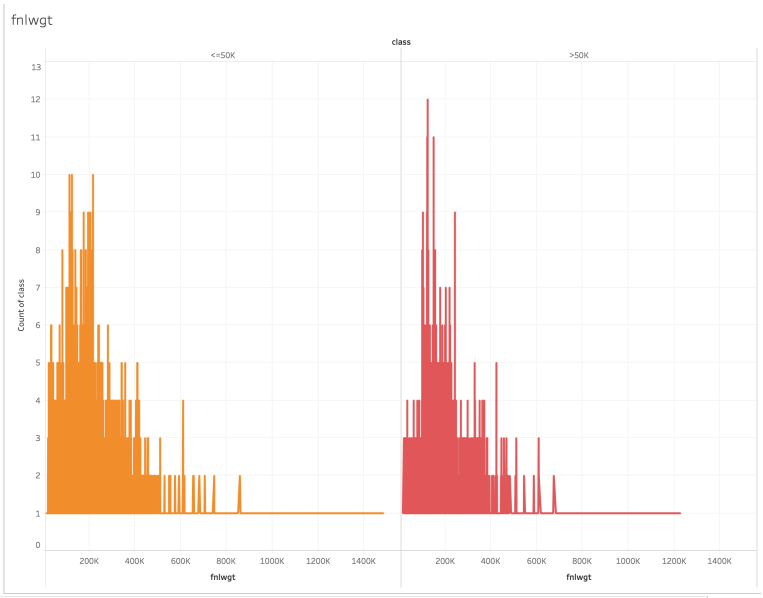
15

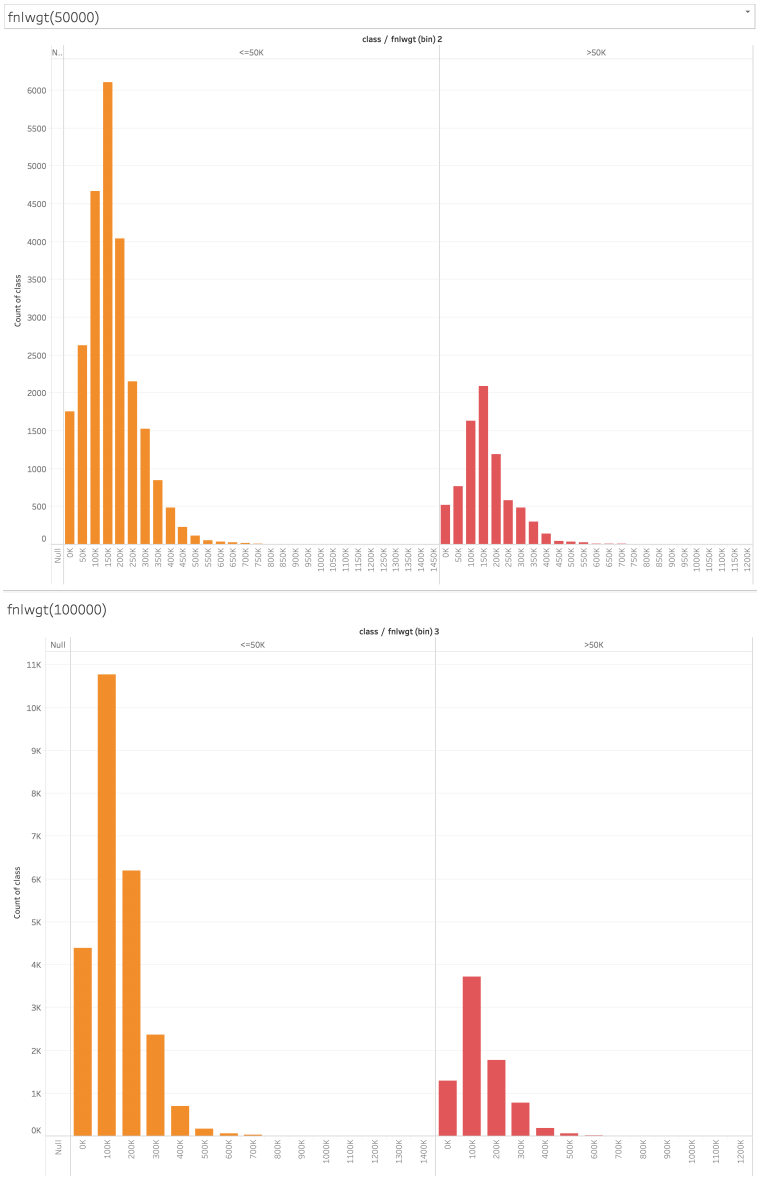


Age: I observe the continuous attribute without discretizing and with discretizing it of equiv-width 5,10,15. The distribution for $\leq 50k$ category follows exponential distribution and for $> 50k$ category follows normal distribution with light positive skewed. (Strong correlation).

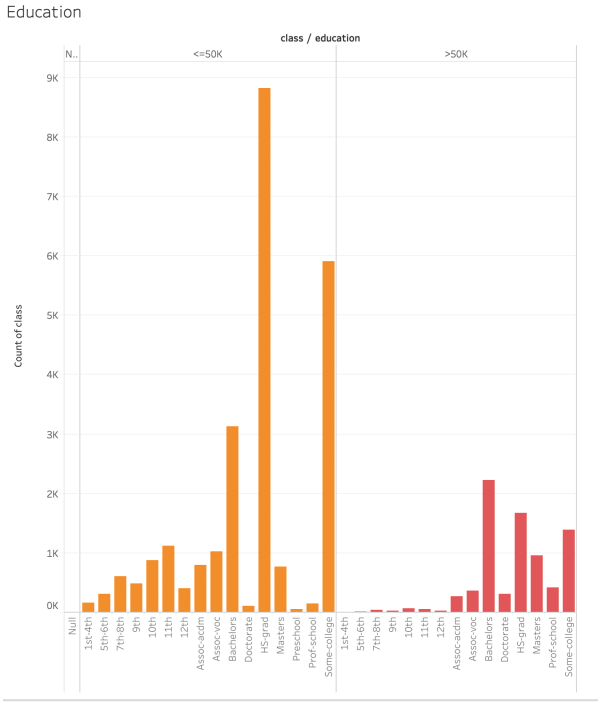


Workclass:More than two third of the work class is private, we can assume that most of the missing value is private in work class attribute. (Weak correlation).

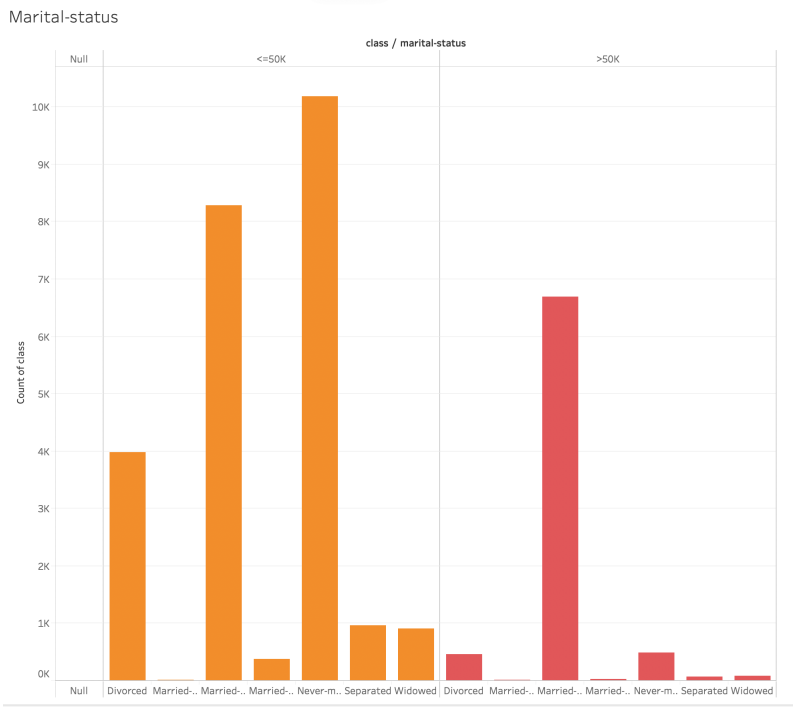

fnlwgt
class
<=50K
>50K
Count of class
fnlwgt
fnlwgt(20000)
class / fnlwgt (bin)



Fnlwgt: I observe the continuous attribute without discretizing and with discretizing it of equiv-width 20000,50000,100000. The distribution of both graphs are very similar. Therefore, it means the fnlwgt attribute don't really play a role on predicting dependent variables. (Weak correlation)

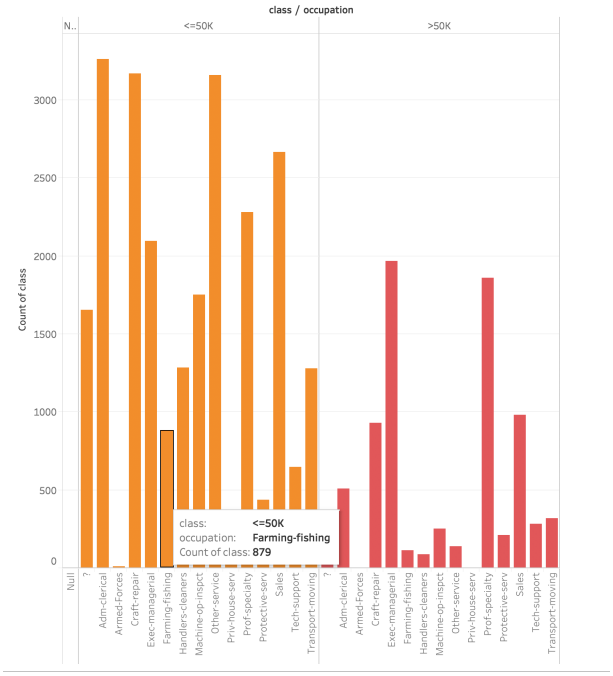


Education :Education and Education-num are the same. Therefore, we can only observe education attribute. People with higher education such as Doc, Master, Pro-school are most likely to have $> 50k$. People with other education are most likely to have $\leq 50k$. (Medium correlation)

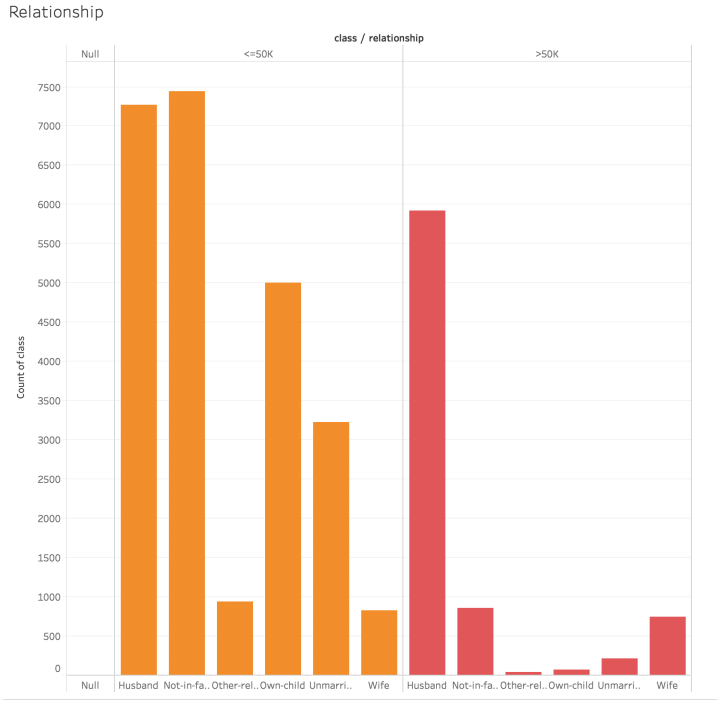


Marital-status: People with Married-civ-spouse is most likely to have $> 50k$. People with Divorced and Never-married are most likely to have $\leq 50k$. (Strong correlation)

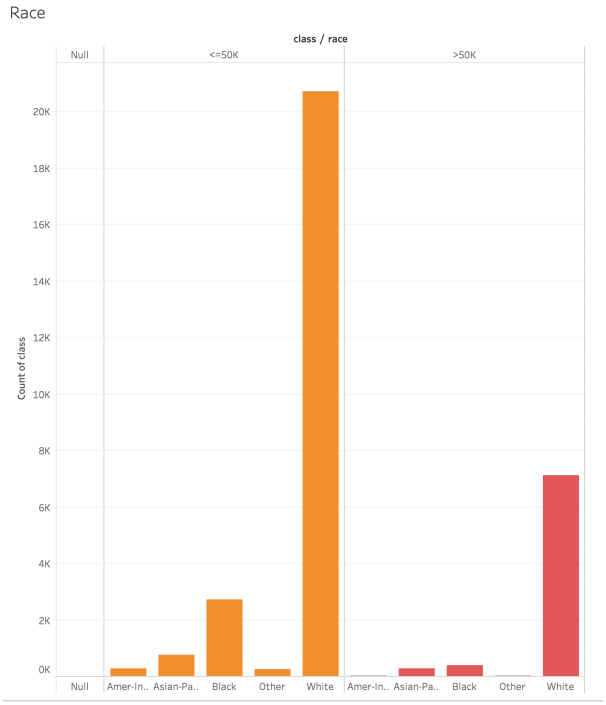
Occuupation



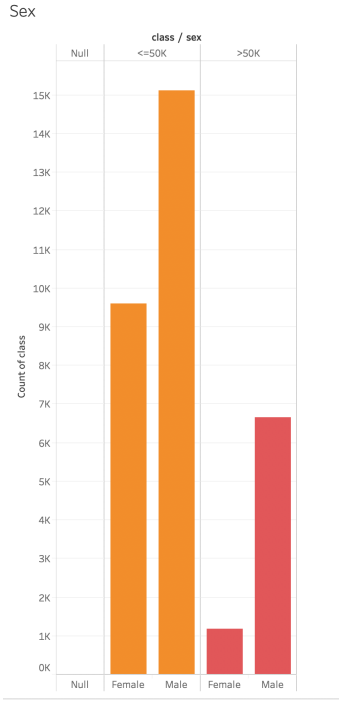
Occupation: The occupation for $\leq 50k$ is almost evenly distributed while Exec-managerial and Prof-specialty have a higher chance falling into the category $> 50k$. (Strong correlation)



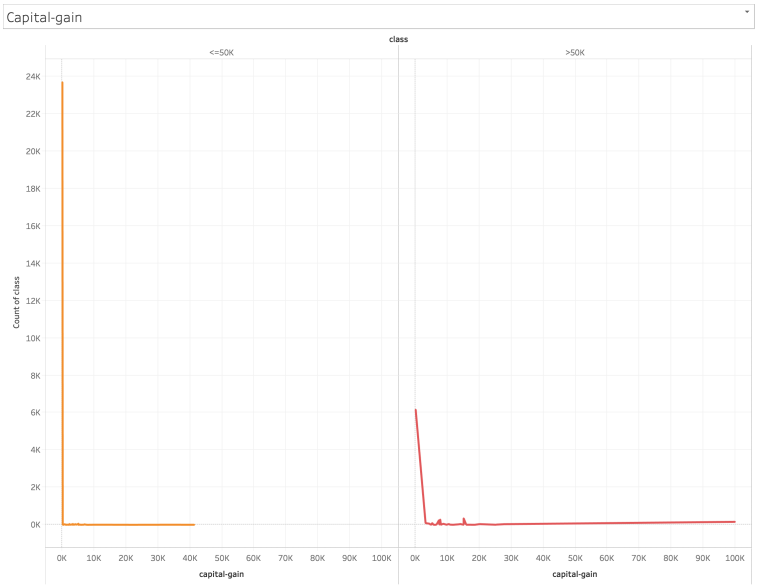
Relationship: For Relationship, husband has a higher chance to fall into the category $> 50k$ than other relationships. (Strong correlation)



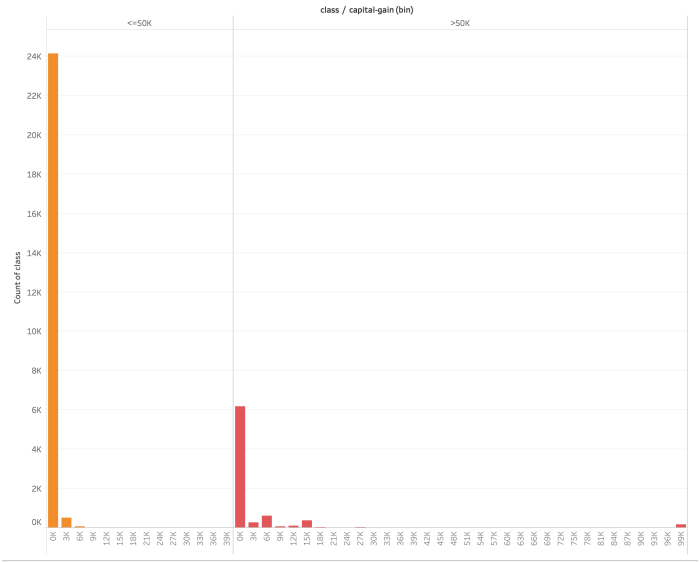
Race: For race, black has a higher chance to fall into the category $\leq 50k$ than other race. (Medium correlation)



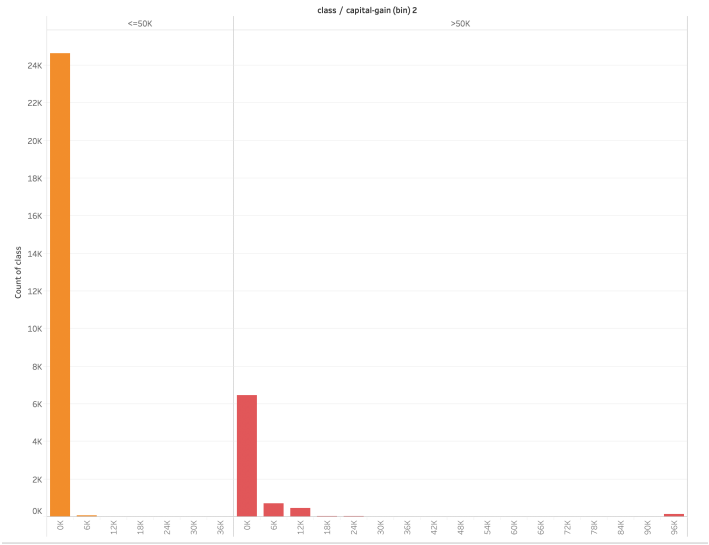
Sex: There 6662+15128 = 21790 male and 1179+9592 = 10771. It was not evenly distributed between two category. Yet we still can see that female has a higher chance to fall into the category $\leq 50k$. Compare 9592/24720 = 0.388 and 1179/7841 = 0.25. Male has a higher chance to fall into the category $> 50k$. Compare 15128/24720=0.612 and 6662/7841 = 0.85. (Medium correlation)



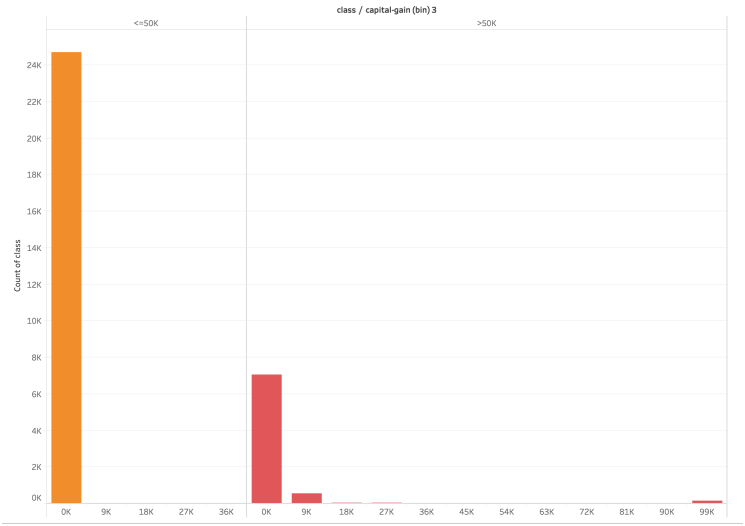
Capital-gain3000



Capital-gain6000

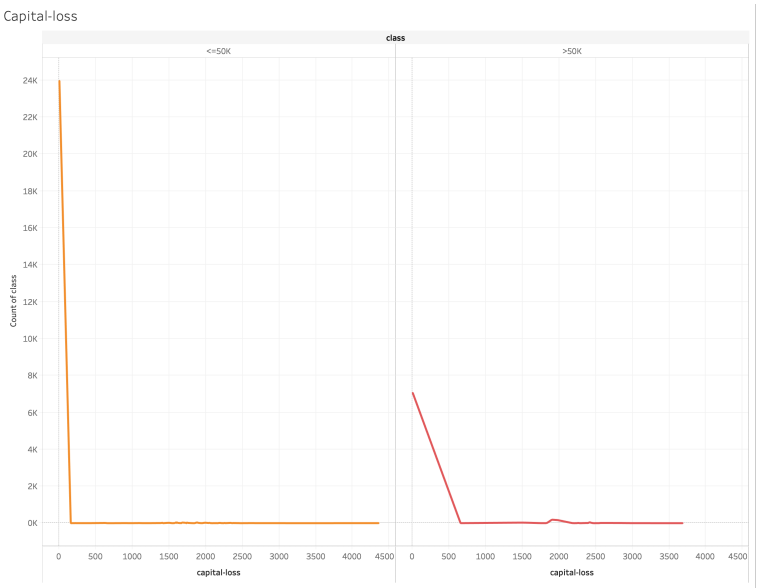


Capital-gain9000

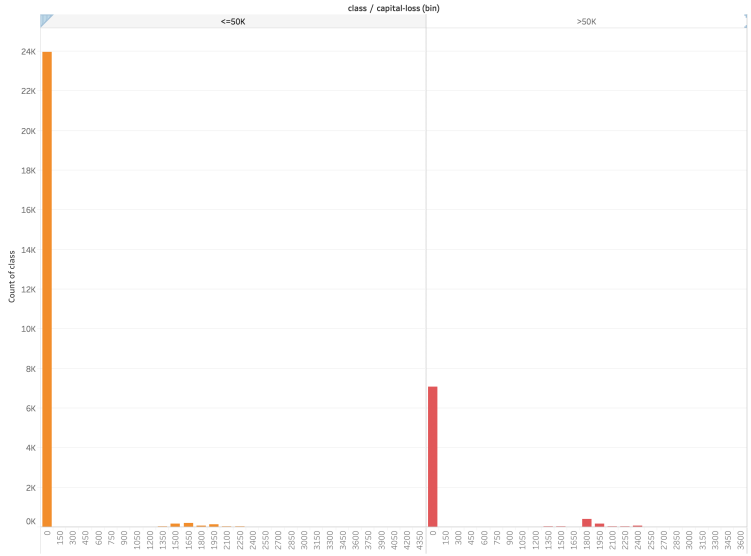


Captial-Gain : I observe the continuous attribute without discretizing and with discretizing it of equiv-width 3000, 6000, 9000. People with higher than 5000 capital-gain are most likely to have $> 50k$. (Medium correlation)

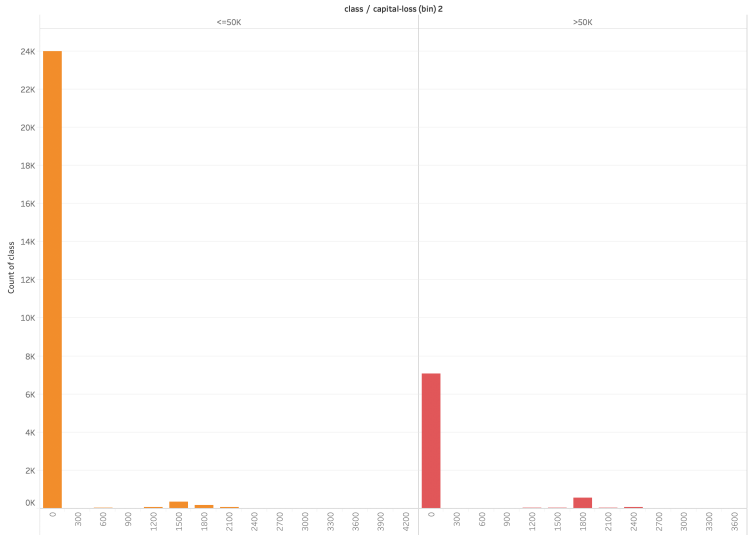
Capital-loss



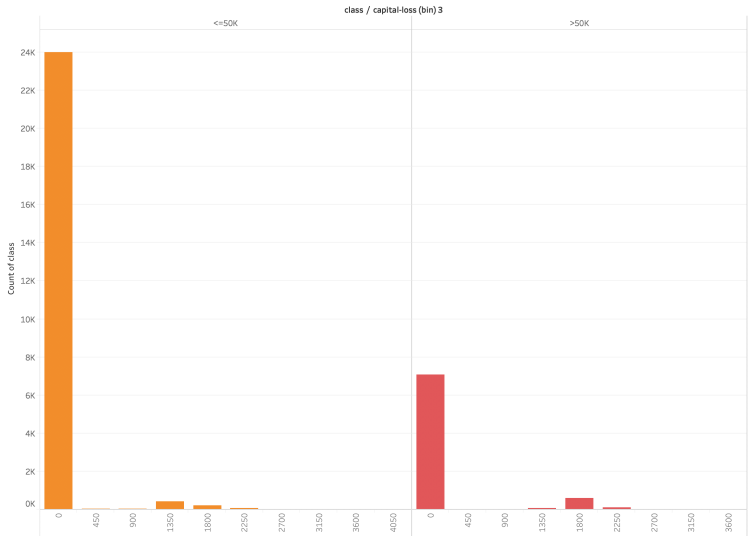
Capital-loss150



Capital-loss300

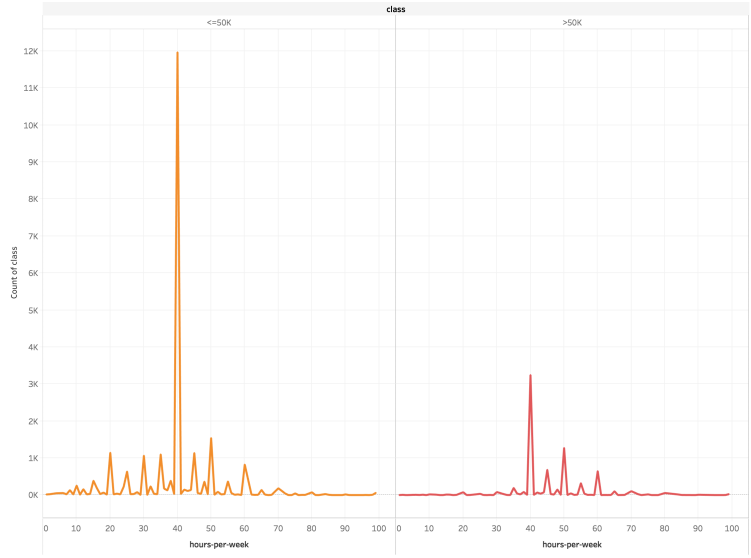


Capital-loss450

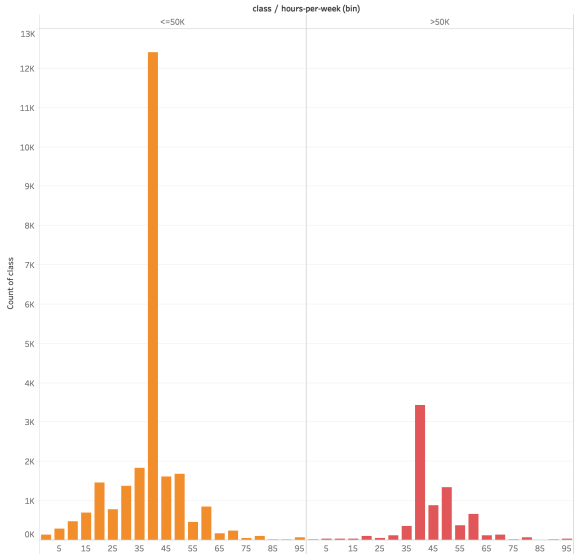


Captial-Loss: I observe the continuous attribute without discretizing and with discretizing it of equiv-width 150, 300, 450. No directly correlation observed in capital-loss.. (Weak correlation)

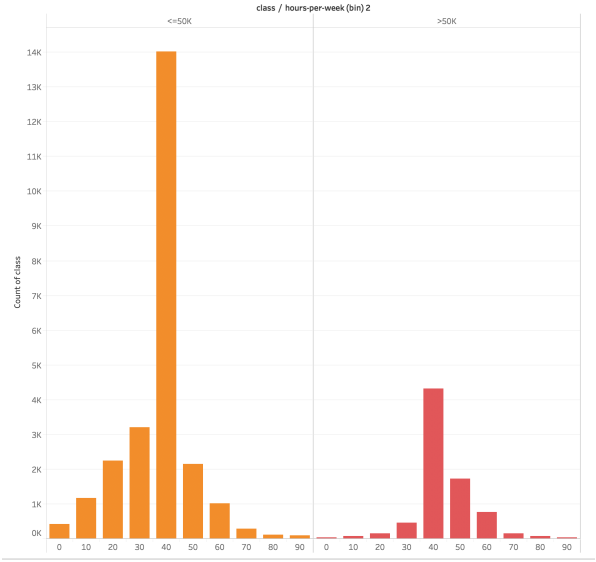
hours-perweek



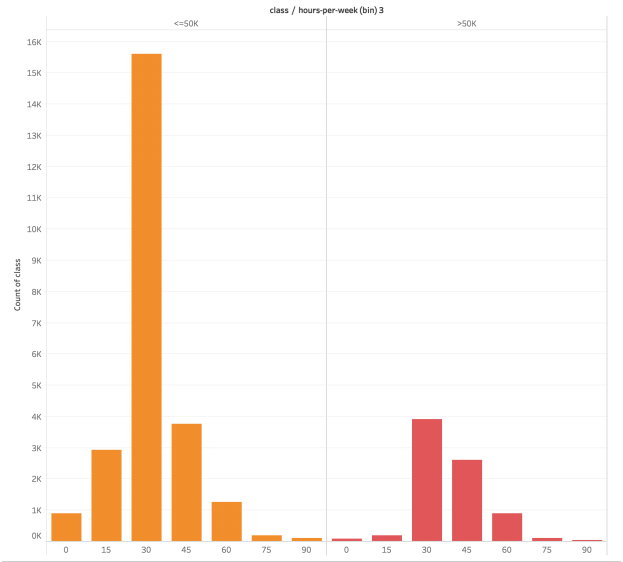
hours-perweek5



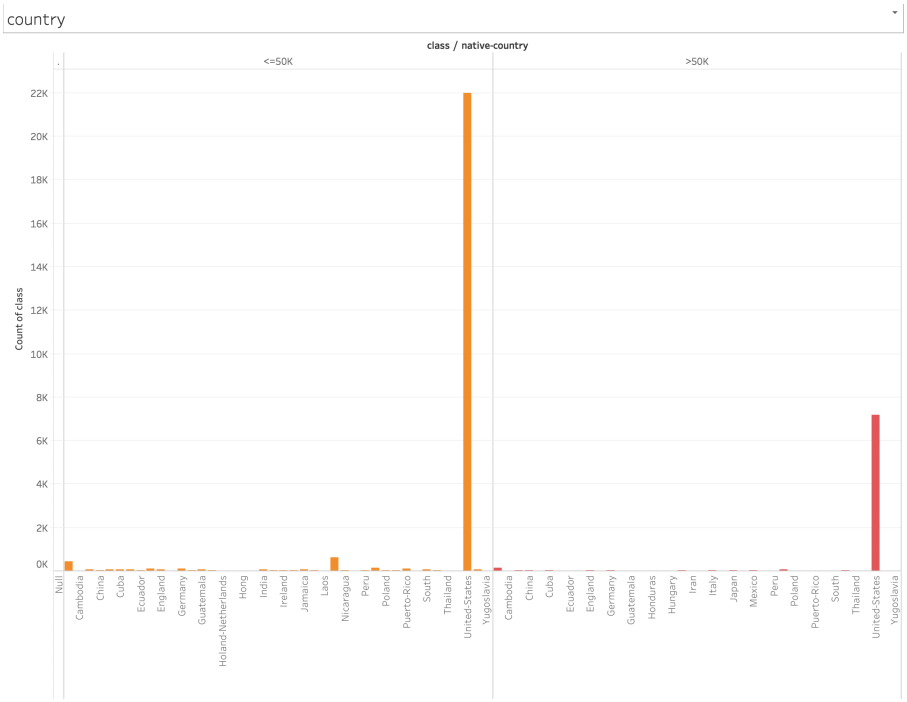
hours-perweek10



hours-perweek15



Hours-Perweek: I observe the continuous attribute without discretizing and with discretizing it of equiv-width 5, 10, 15. People work less than 40 hours per week are tend to fall into the $\leq 50k$ category. (Medium correlation)



Country : Most people's native-country are united-state. 21999+7171 = 29170, 29170/32560 = 0.895. We can assume the missing value for attribute native-country are united-state. (Weak correlation)