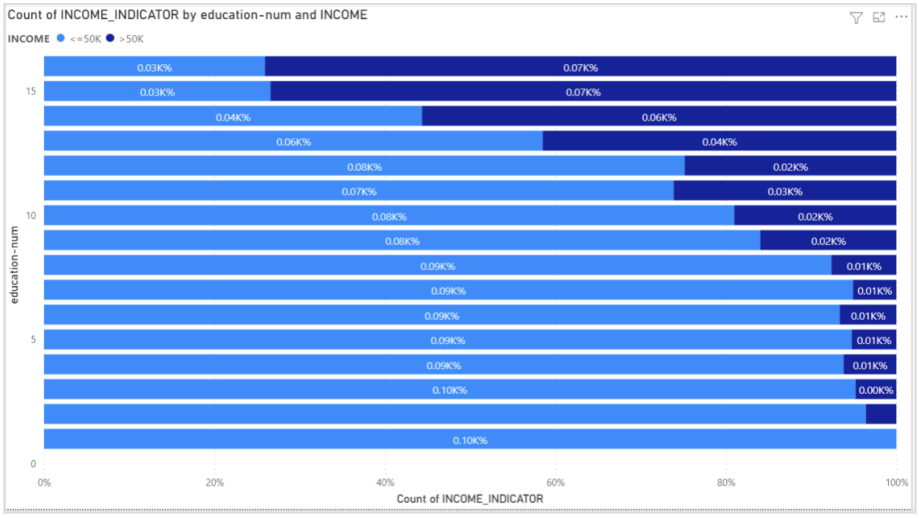
System Documentation Report

Project Statement:
UVW college provides a dataset to be analyzed to check which attributes affect the salary (>50k, or <=50k). We as XYZ's analysis team, analyzed the data and based on our results we have compiled this report**.**

Roles & Responsibilities:
Throughout the team meeting, we analyzed the dataset and then divided the attributes among team members. Each member visualized and analyzed the assigned features. Then we share our thoughts and interpretations together to get final insights and decisions.

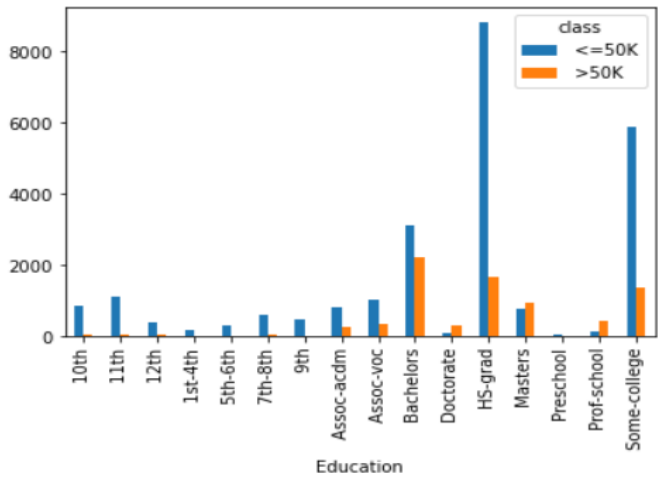
Name	Role
Goabaone Ramolefe	education num,age, marital status
Bingxi Li	capital-gain, relationship, and workclass
Islam Hamed	capital-loss, occupation, country-native, and sex
Mohammad Jaradat	education, race, hours-per-week and workclass
Ahmed Rafaay	hours-per-week, race, and fnlwgt

Education-num: We want to figure out if the education-num of individuals affects their income label (>50k, or <=50k) and whether it can be used in the income prediction model.



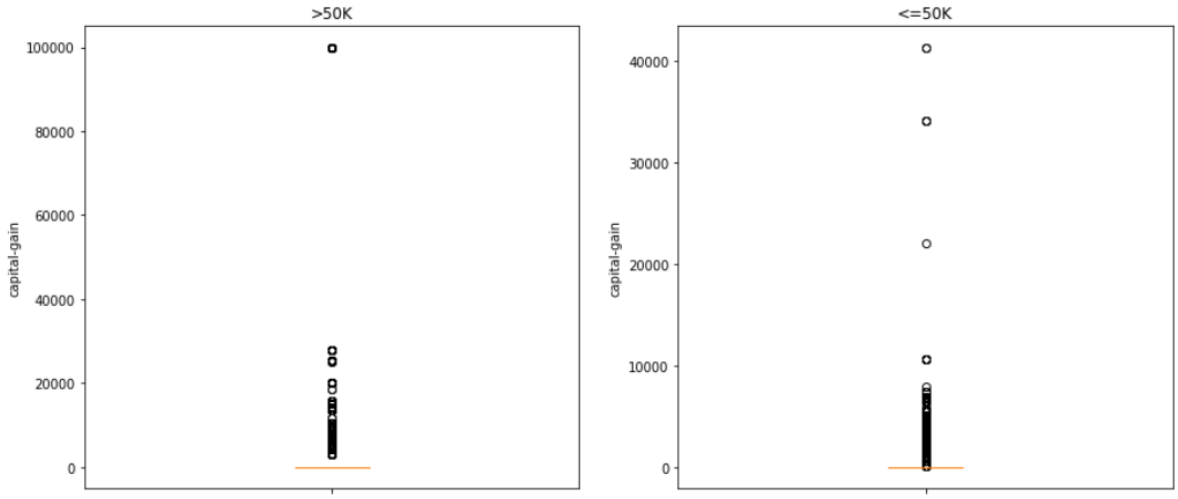
Interpretation : from the above stacked bar chart we can see that education number is an important attribute in determining the income of an individual. Looking at the percentages per education number we noticed that as the education number increases the percentage number of those earning above 50k increases significantly.We concluded that this is an important factor and we included it as the number one factor compared to other attributes.

Education: We want to figure out if the Education of individuals is affecting their income label (>50k, or <50k).



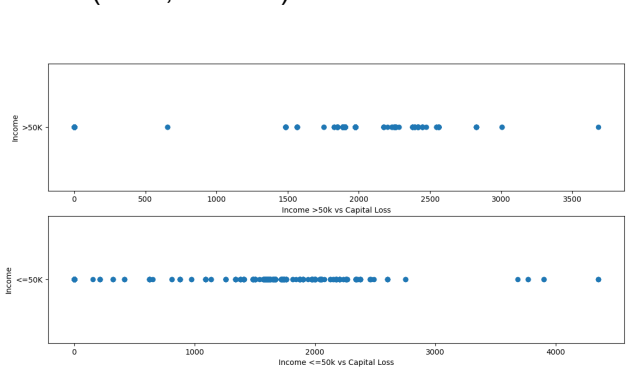
Interpretation: the above unstacked bar chart shows the individuals whose salaries are categorized within two classes (>50k, < 50k). In this chart, we showed the data based on one attribute which is education. Moreover, we can see that three options (Prof-School, Masters and Doctorate) are more likely to gain a salary above 50k (>50k) with probability of more than 60%, especially for the Prof-School individuals. However, all the other options from this attribute (Education) are more probable to have a salary less than 50K (<50k).

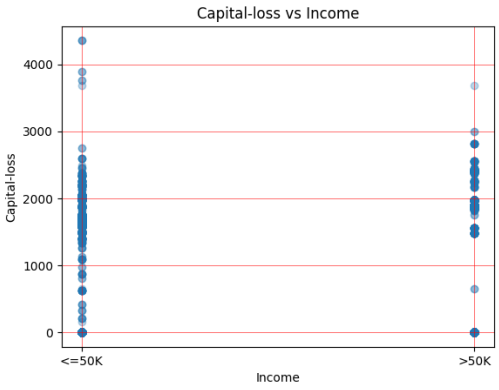
Capital-gain: We want to figure out if the Capital-gain of individuals affects their income label (>50k, or <=50k) and whether it can be used in the income prediction model.



Interpretation: the boxplots show the distribution of capital gains for the group with income >50k and the group with income <=50k. With the median capital gain of both groups being 0, this feature alone is probably not a sufficient salary group indicator for more than half of the individuals with no or low capital gains. However, as the majority in the salary <=50k group have capital gains less than 10k (more than 99.9%), it can be concluded that if an individual has capital gains more than 10k, he or she is almost certainly in the income >50k group. We therefore decided to include this feature as a relevant factor, which can be a very strong and accurate indicator for high salary prediction and serves the overall goal of income labelling when combined with other factors.

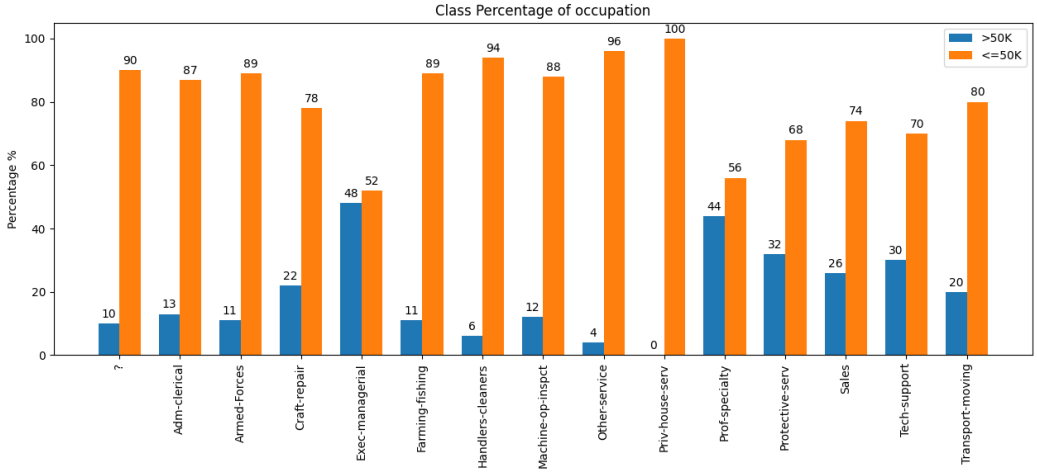
Capital-loss: We want to figure out if the capital-loss of individuals is affecting their income label (>50k, or <50k).





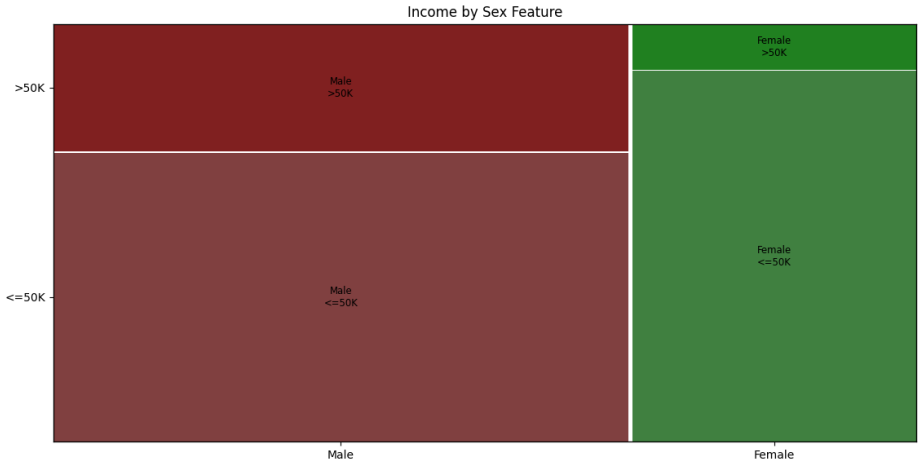
Interpretation: the chart provides us with an insight of capital loss against income. We can see the capital range where they have both labels. However, there is another range that classifies one of them. For example, the range from 2800:3800 goes to >50k class. For 1:1400 goes to <=50k. So, if an individual's capital loss is 3000 he is absolutely getting >50k.

<u>Occupation</u>: We want to figure out if the occupation of individuals is affecting their income label (>50k, or <50k).



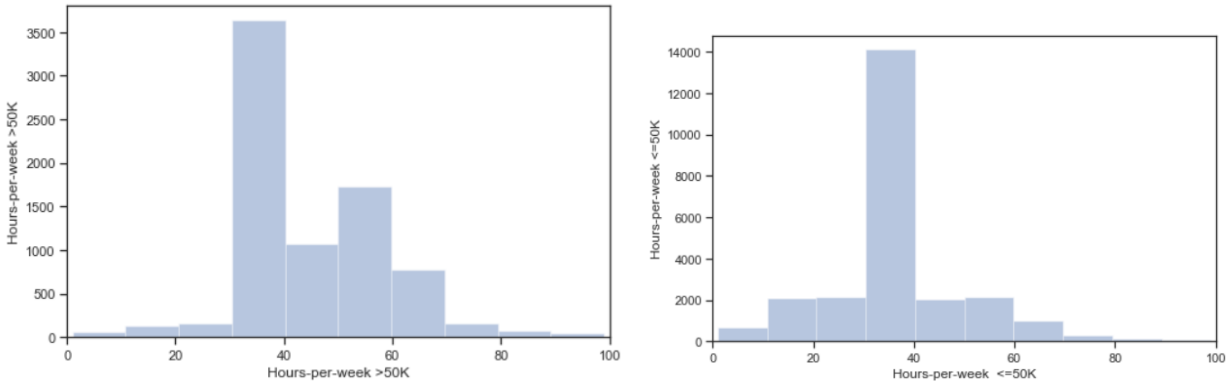
Interpretation: the chart provides us with insights of occupation against income. The data is counting the number of individuals in each category (occupation), then calculating the percentage of income label in each occupation. We can see that a higher percentage (>50%) is with income <=50k. So, this feature cannot classify both income classes (>50k, and <=50k). However, this feature can be used to indicate for class <=50k by percentage. For instance, if the individual's occupation is priv-house-serv, so definitely his/her income is <=50k. According to the dataset, every individual' occupation is priv-house-serv her/his income <=50k, as shown in the chart. Also, we can use it beside other features as supporting decisions by percentage.

<u>Sex</u>: We want to figure out if the sex of individuals is affecting their income label (>50k, or <50k).



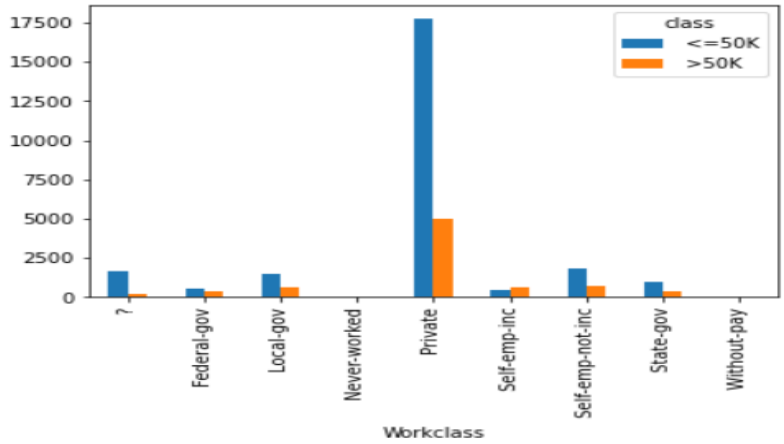
Interpretation: the chart provides us with an insight of gender against income. The data is counting the number of individuals in each category (male, female), then calculating the percentage of income label in each one. We can see that a higher percentage (>50%) is with income <=50k. So, this feature cannot classify both income classes (>50k, and <=50k). *However*, we can interpret that male is more likely to have an income >50k than females despite the number of male in the dataset is greater than the number of females as shown in the Mosaic chart.

Hours-per-week : We want to figure out if the Hours-per-week of individuals is affecting their income label (>50k, or <50k).



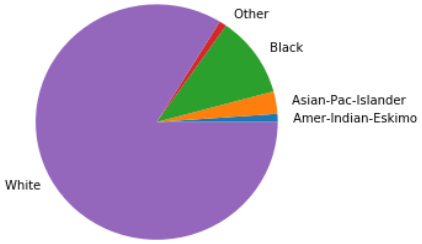
Interpretation: Based on the provided charts, we can see that it's very hard to determine whether the individual could gain a salary above 50K (>50k) or not based on the Hours-per-week attribute. The same as in Race, we can witness the same distribution among the individuals for both classes, which results in a fair classification, if we need to rely on this attribute.

Work-class : We want to figure out if the Work-Class of individuals is affecting their income label (>50k, or <50k).

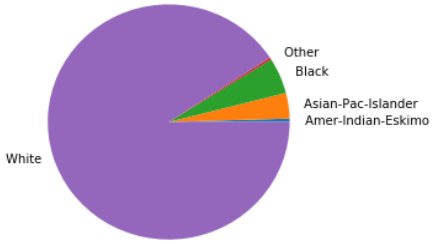


Interpretation: From the above mosaic and bar charts, we can see that this attribute also can not be used to determine the individual's income. In addition to that, this attribute has not shown any relationship with salary above 50K (>50k), so this will not be helpful and even will have drawbacks if we use it in our overall analysis due to not relevancy with our goal. Finally, the same as some other attributes, and based on this attribute performance, we exclude it from the executive report.

Race : We want to figure out if the race of individuals is affecting their income label (>50k, or <50k).



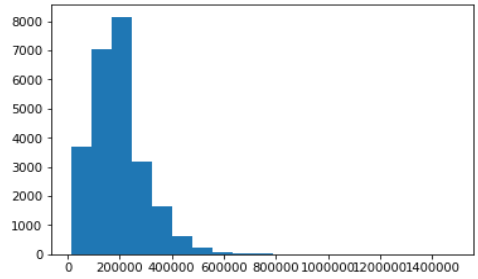
Salary >50K



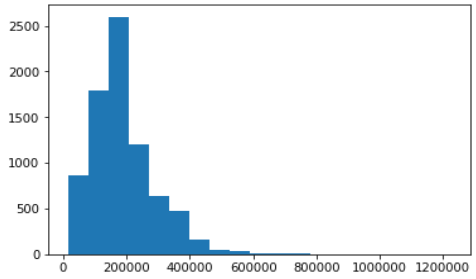
Salary <= 50K

Interpretation: From the above pie charts, we can see that this attribute also can not be used to determine the individual's income. These diagrams show that the survey was mainly targeted towards White people. The distribution is nearly the same for both low and high-income people. As a result, this feature won't be useful.

Fnlwgt : We want to figure out if the final weight of the survey sample of individuals is affecting their income label (>50k, or <50k).



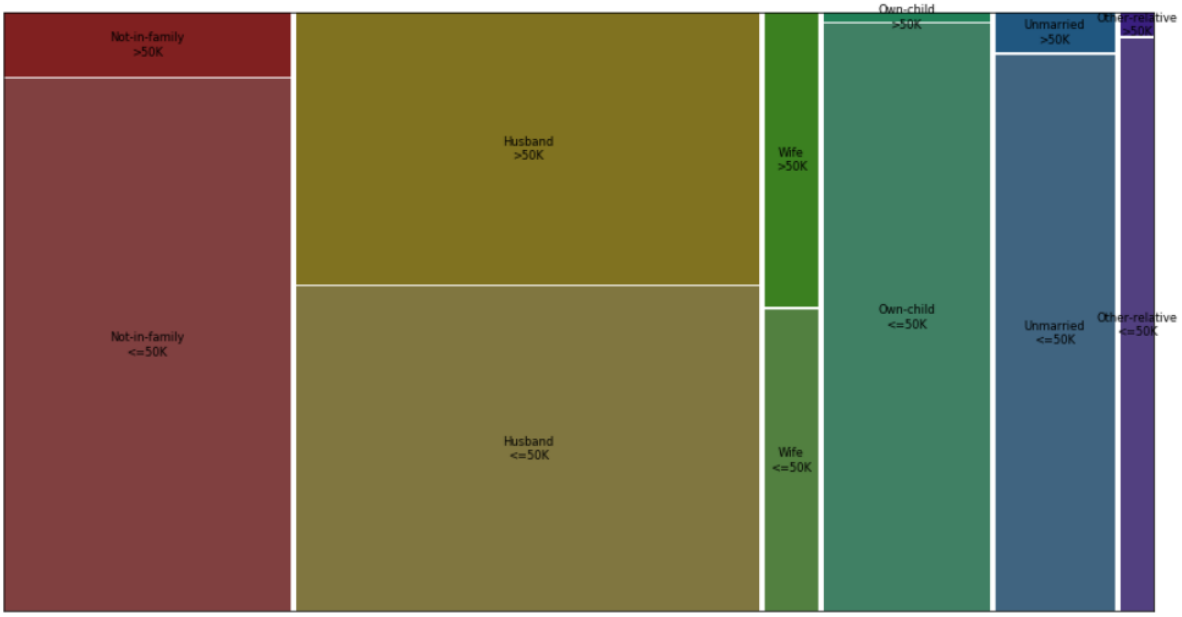
Salary <= 50K



Salary >50K

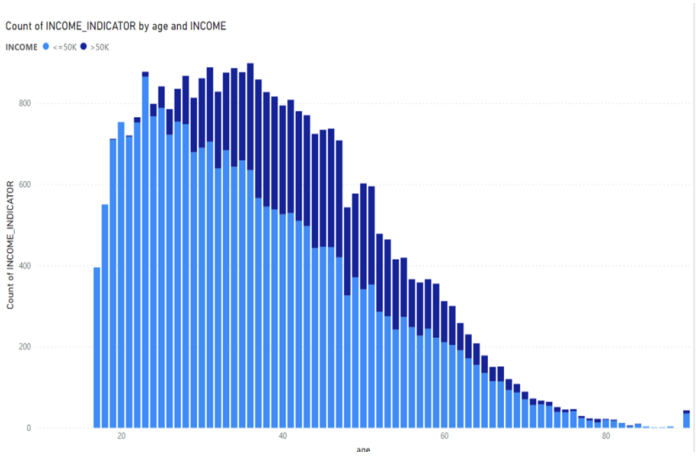
Interpretation: From the above bar charts, we can see that this attribute also can not be used to determine the individual's income. We nearly have a similar histogram for both income levels. This reason, in addition to the low percentage of respondents in the second group, make this feature not useful for salary determination.

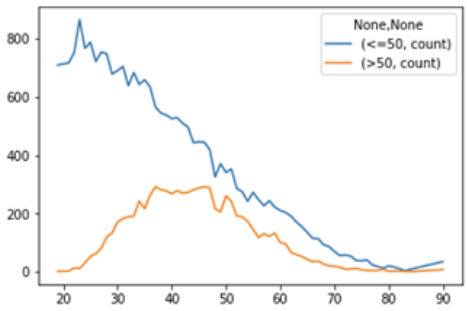
<u>Relationship</u>: We want to figure out if relationship of individuals affect their income label (>50k, or <=50k) and whether it can be used in the income prediction model



Interpretation: the mosaic graph shows the distribution of income <=50k and >50k per each relationship cluster. Individuals identified as Husband or Wife have the probability of around 50% to earn more than 50k (46% and 49% respectively), while individuals in other relationship groups are less likely to earn more than 50k. We concluded that the factor relationship is relevant to income. However, we didn't include it as the top four factors as for the majority of the observations (Husband and Wife), the factor does not indicate a higher probability towards a certain income group and therefore can not be used for labeling.

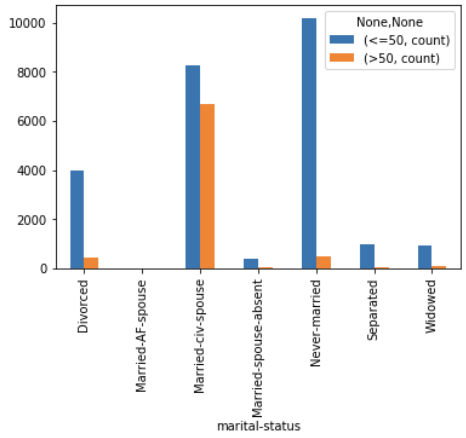
<u>Age</u>: We want to figure out if the education-num of individuals affects their income label (>50k, or <=50k) and whether it can be used in the income prediction model.

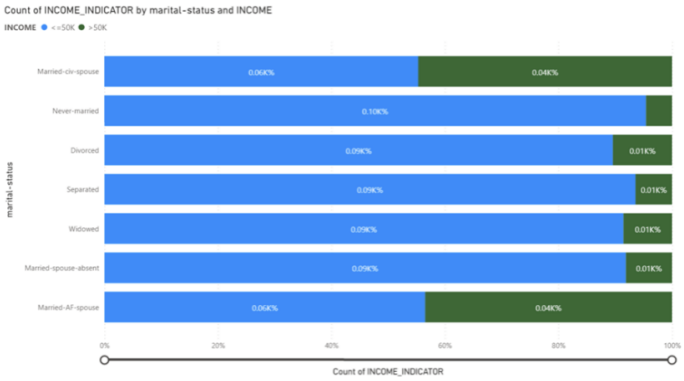




Interpretation: The graphs show that as age increases the numbers of people earning either above or below 50k decreases. We noticed that although the number of people earning 50k increases at some point ,the general trend is that as the years increases there is always more that 50% of individuals who are earning less than 50k compared to those that are earning more than 50k and based on that we concluded that age was not a good measure to predict income.

<u>Marital-status</u>: We want to figure out if the marital-status of individuals affects their income label (>50k, or <=50k) and whether it can be used in the income prediction model.





Interpretation: the visualisations above show that marital status is not a good attribute for predicting income. The bar chart on the left shows that for each marital status there is more 90% of people who are earning less than less than 50k and 10% earning more than 50k.We only see one instance under married -civ-spouse where there is almost an equal number from both classes.

Conclusion: The main goal was to determine which attributes have the main effect on class labels (>50k, or <=50k). So, we selected the attributes based on their visualizations, to see how much these attributes can be helpful in classifying both labels in the same visualization.
Based on our analysis and understanding for all attributes and visualizations, we selected the top four attributes based on their performance, which are (Education-num, Education, Capital-gain, and Capital-loss). We worked on selecting attributes that classify the income as >50k, or <=50k. So, the top attributes that clearly distinguish both classes have been selected.
In summary, this report describes each attribute with its visualization, and interpretation based on the analysis we did.

Appendix: python's scripts are zipped with this report as reference.