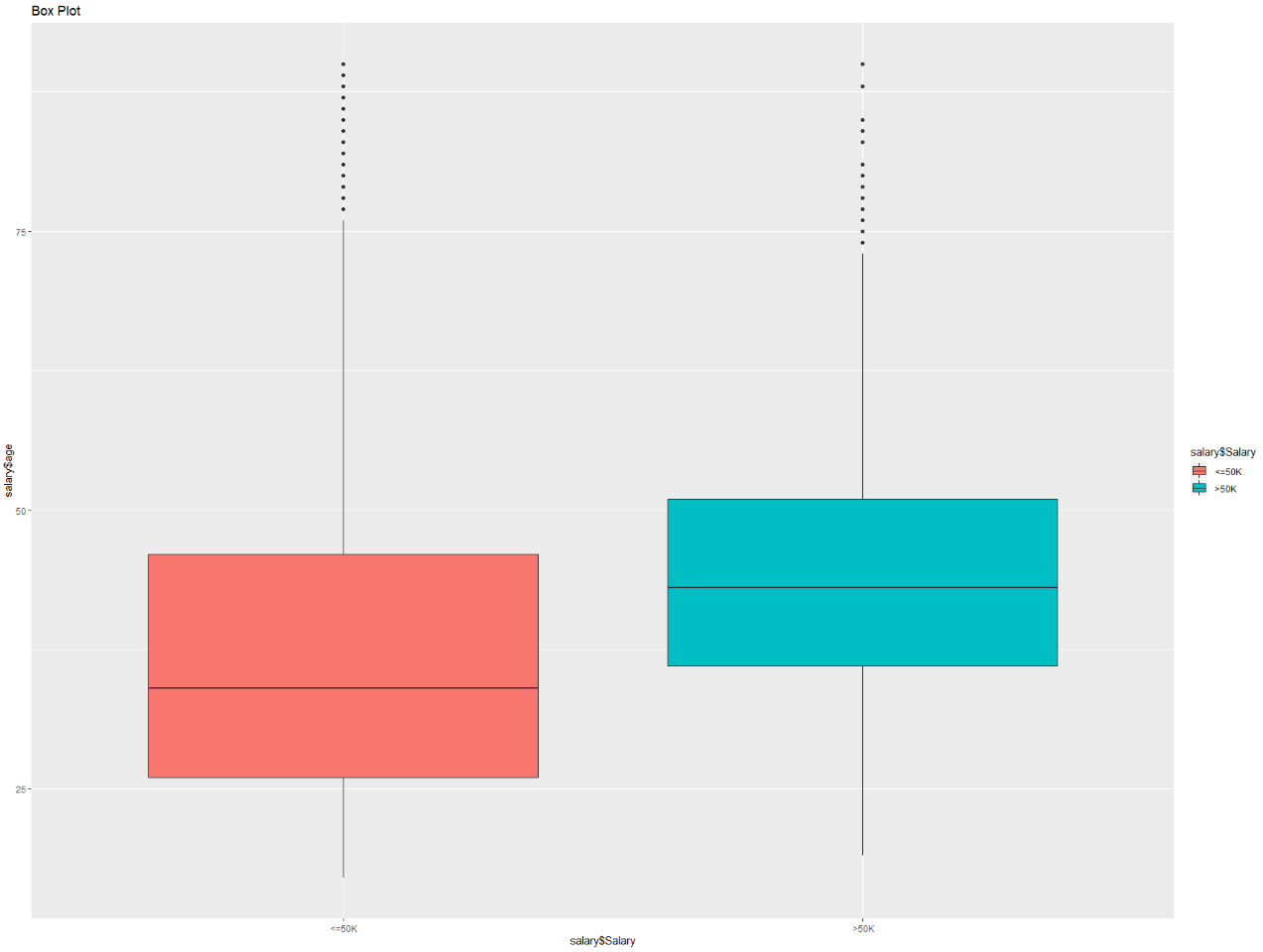
**Naive Bayes**

**Salary Prediction**

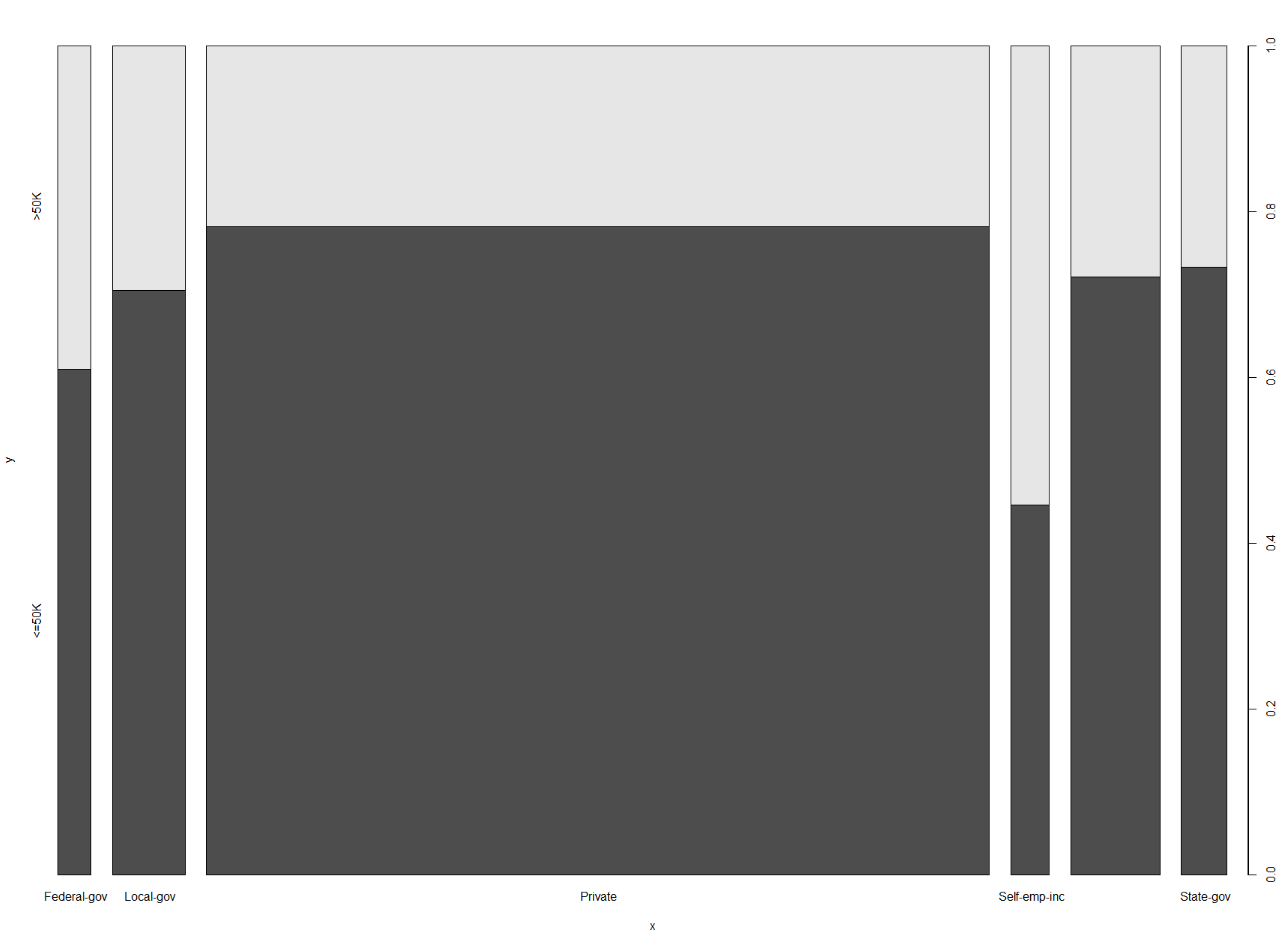
* Since the data is already in factor format, no need to discretize except for education no.
* Performed EDA and got following results:

1. Plot of Age against Salary:



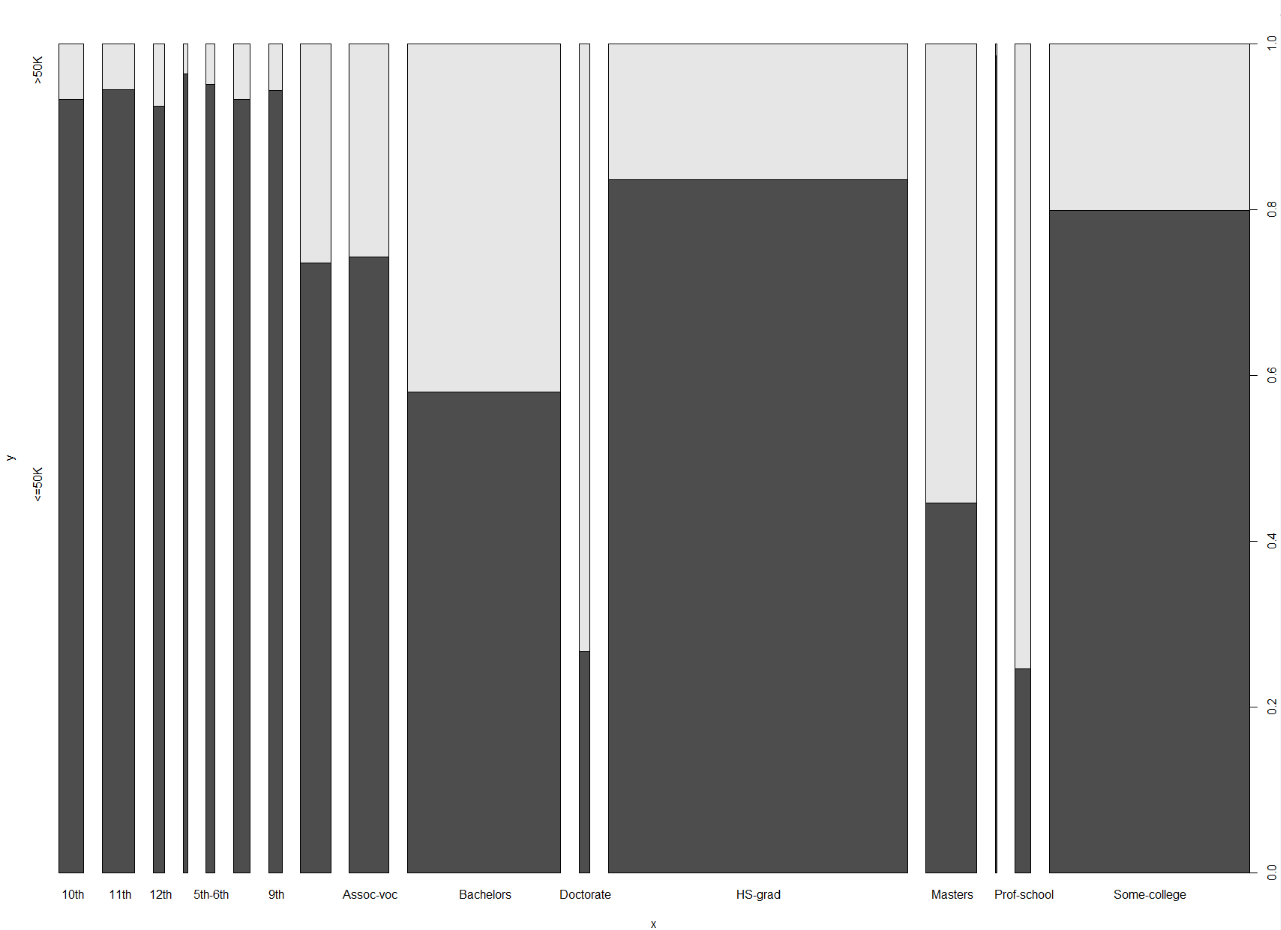
* Observing the plot we can infer that 75% of the people having a salary of above 50k are in their mid-40s or 50s. But there are also young people who have been getting a salary of more than 50k. On contrary there are also older people who are still under 50k, maybe education and other factors might be playing a role in it. We can deep dive by finding the correlation between them.

1. Plot of work class against salary:



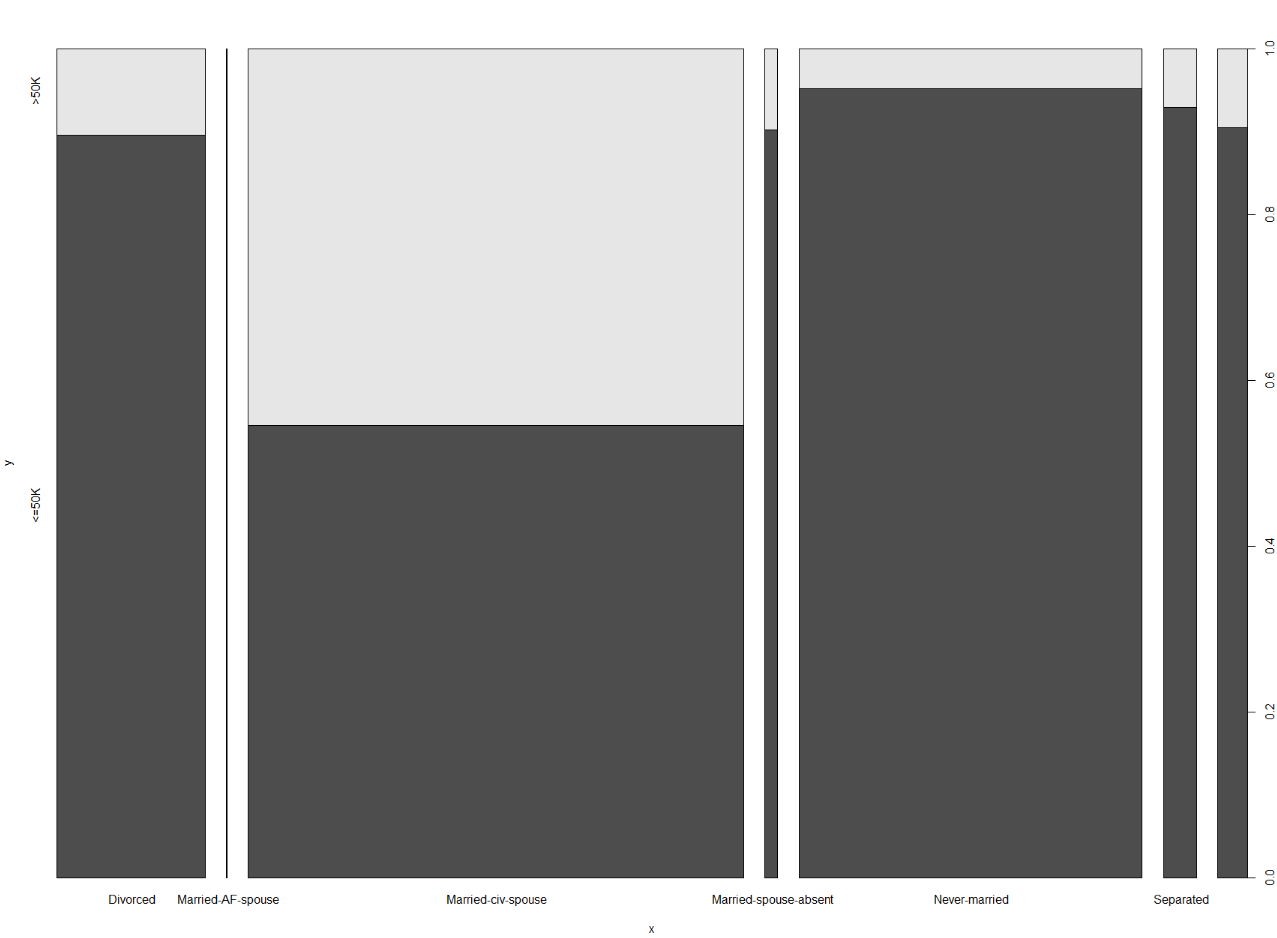
* Observing above plot we can infer that there’s more number of people in the private sector, but almost 80% of them are in the category of less than 50k. While on the other side it’s the self-employed category who has a higher percentage of people in >50k category compared to its own category.

1. Plot of education against salary:



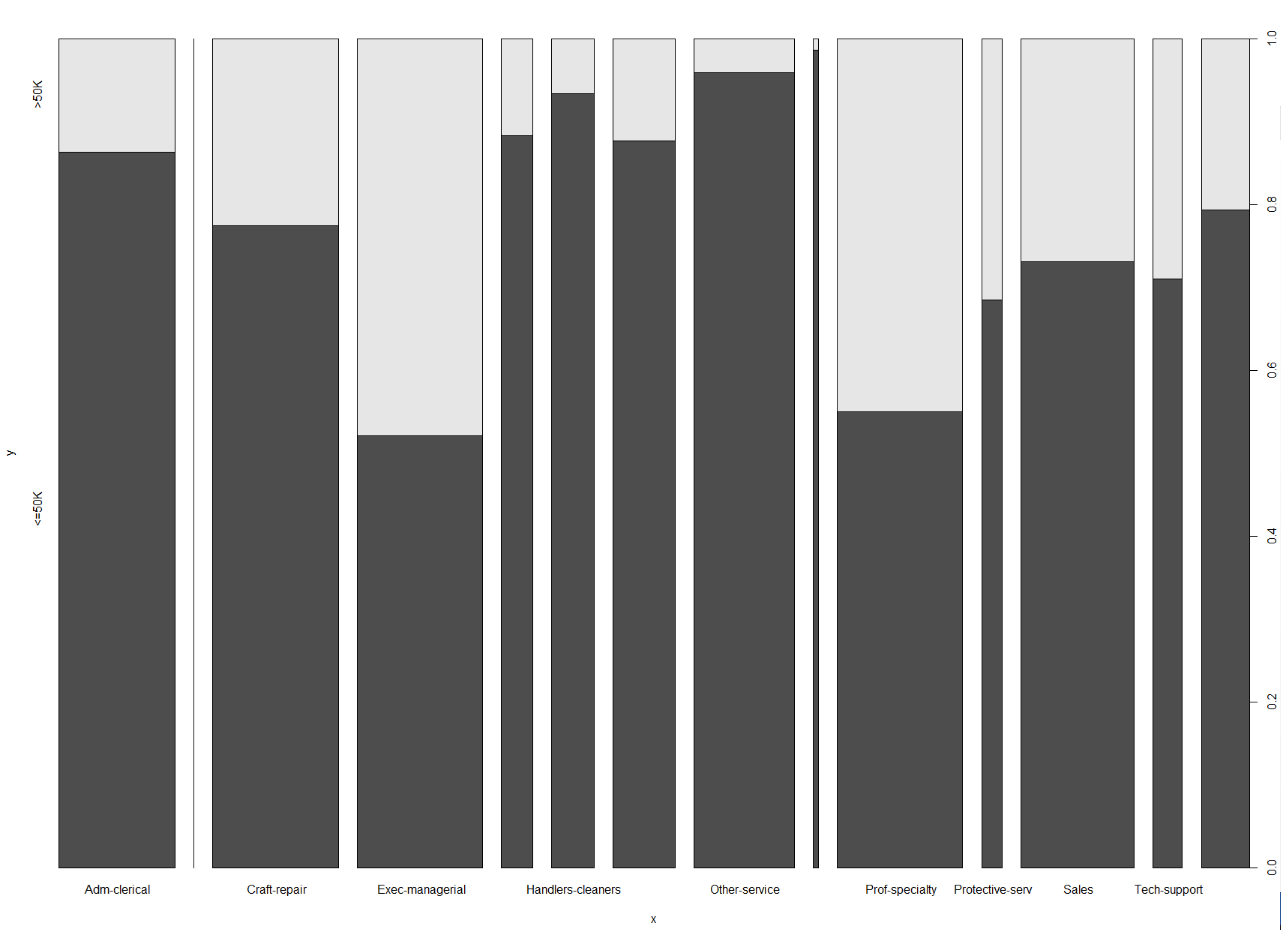
* Observing above plot we can infer that a large number of proportion of people are HS graduated and of those 80% are in the <=50k category. While Doctorate and Prof school category people are less in number but the ones whose around 50% of people have a >50k salary.

1. Plot of marital status against salary:



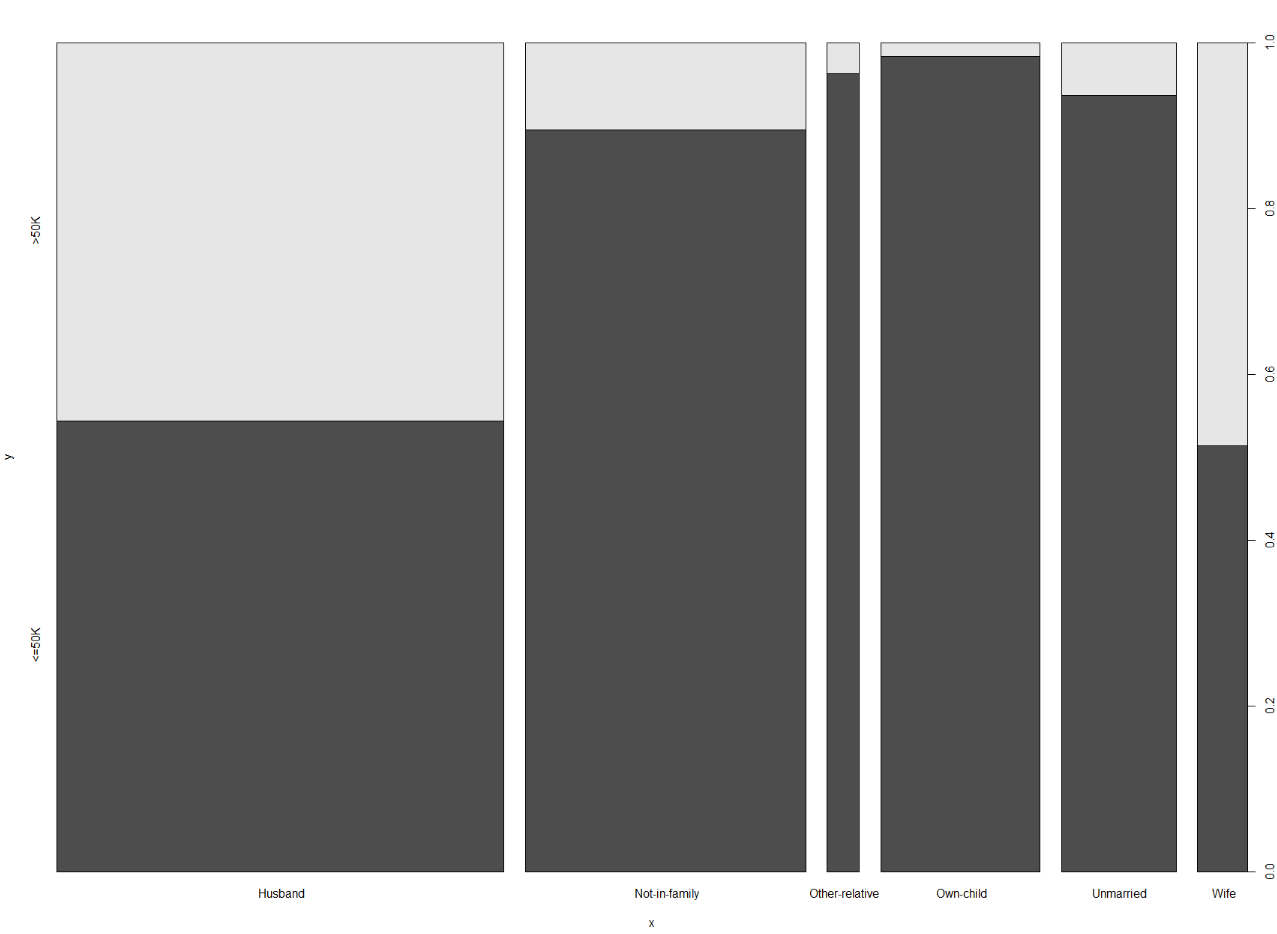
* Observing above plot we can infer that a large of people are married with civilian spouse and of those 50% of people are above 50k. We need to deep dive to see how much parameter is related to the output.

1. Plot of occupation against salary:



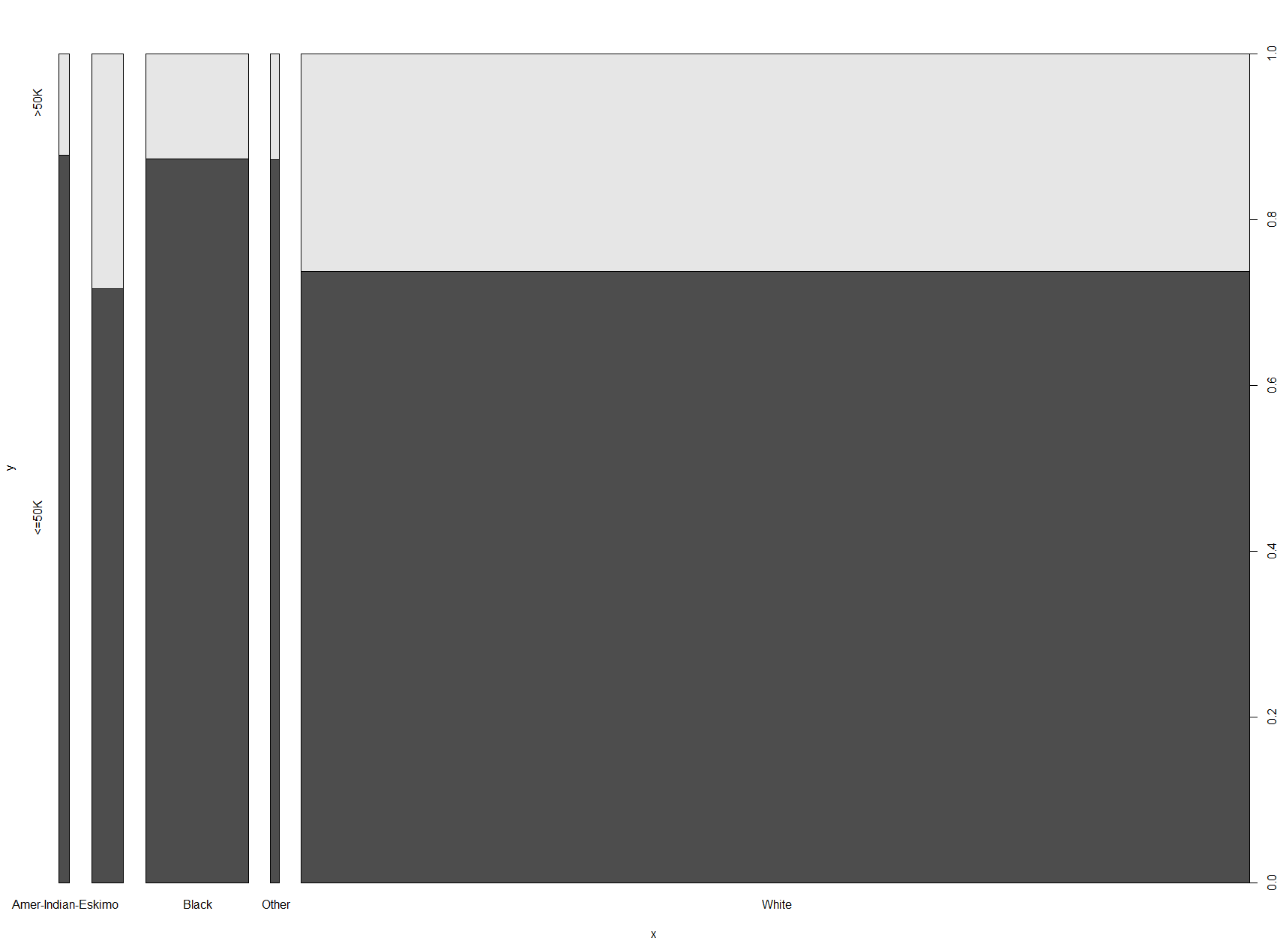
* We can infer from above plot that most of the people earning a more than 50k salary are from Exec-managerial and Prof-specialty.

1. Plot of Relationship against salary:

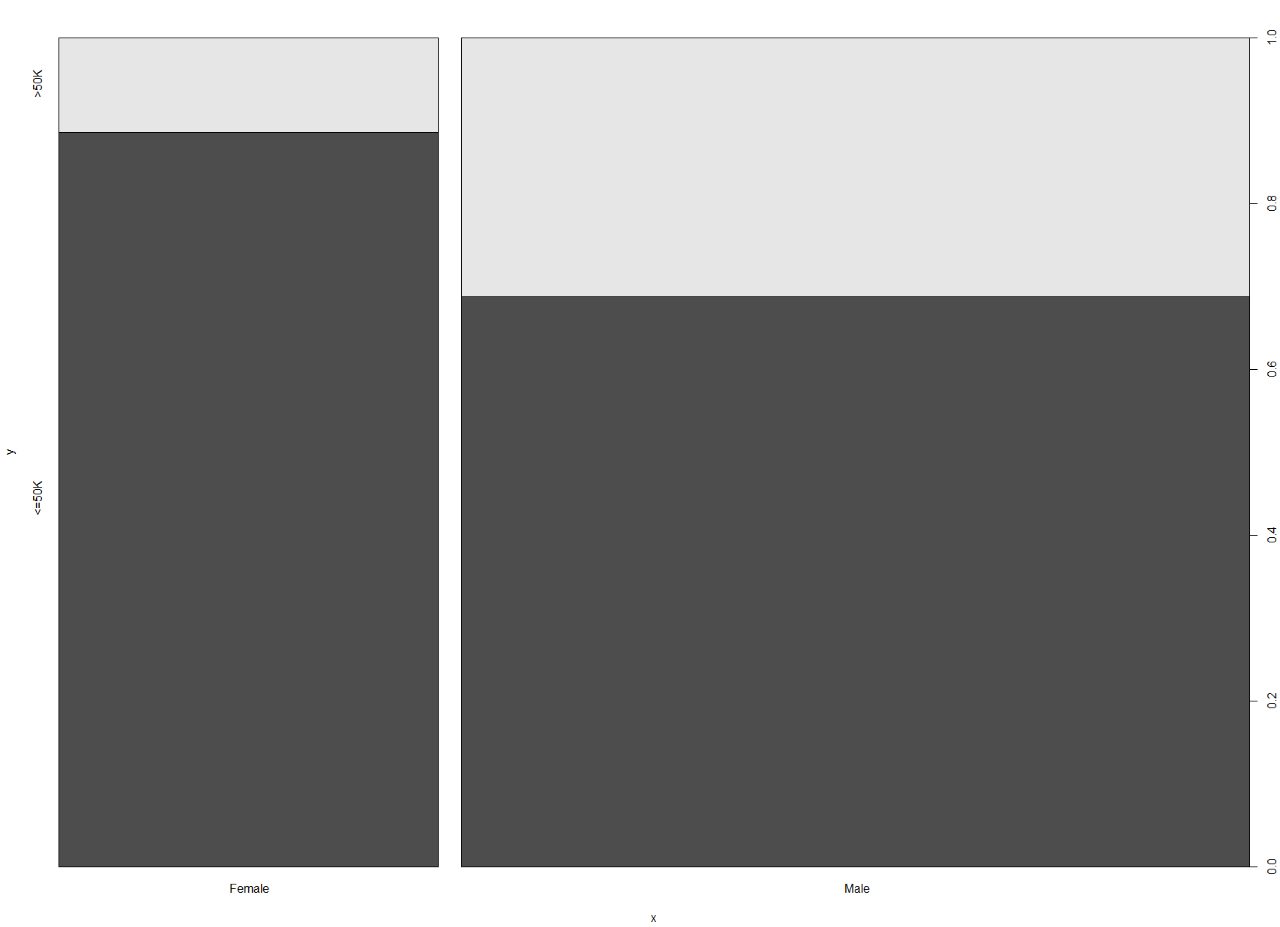


* We can infer from above plot that majority of the people are husband where 50% of people are above 50k. We need to check how this parameter affects the relation with salary.

1. Plot of race against salary:

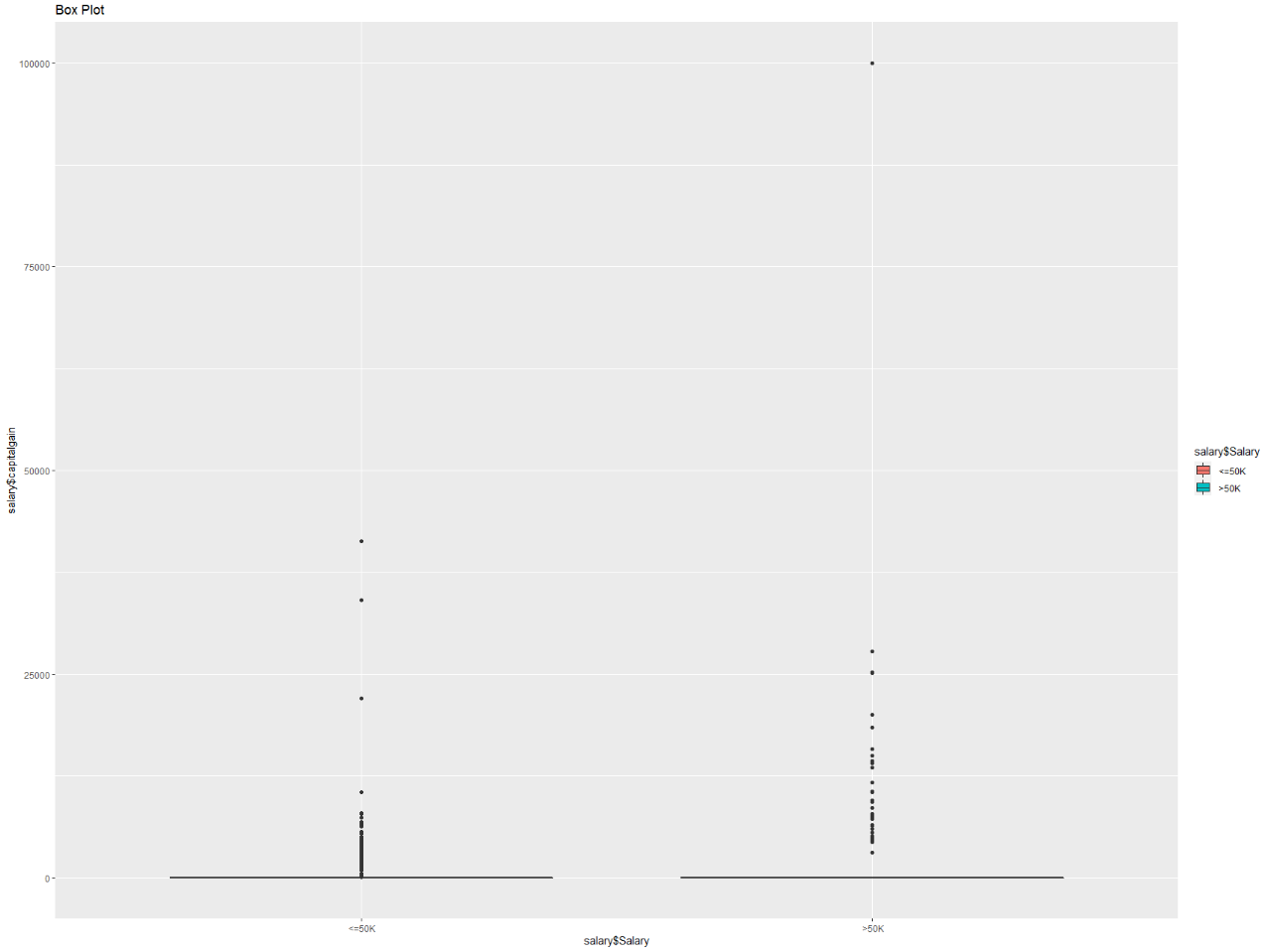


1. Plot of sex against salary:



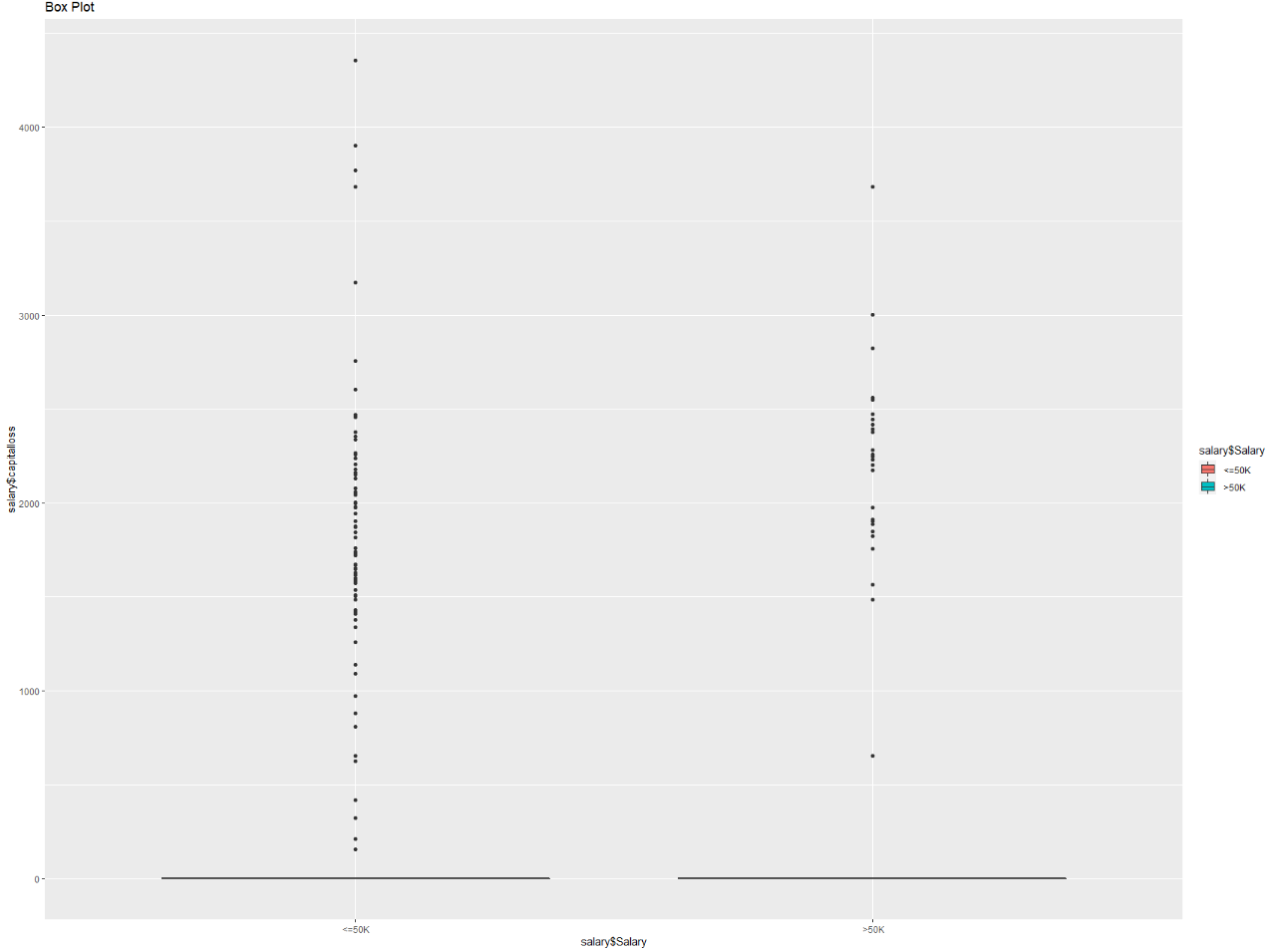
* We can infer from above plot that the male are dominating here, also the proportion of males getting <50k is bigger than of females.

1. Plot of Capital gain and salary:



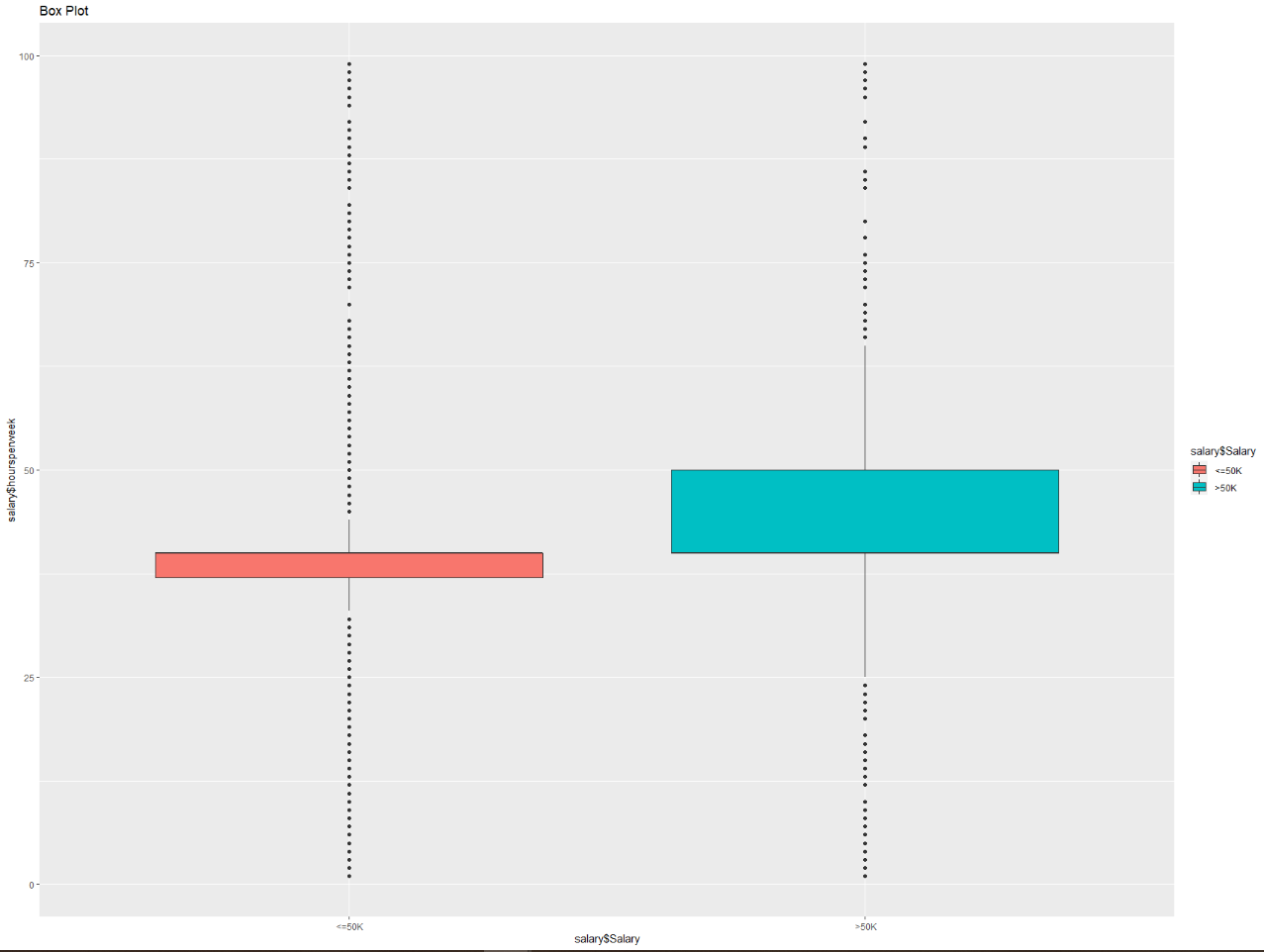
* We can infer from above plot that whether it be <=50k or >50k its only 25% of people getting any capital gain.

1. Plot of Capital Loss against salary:



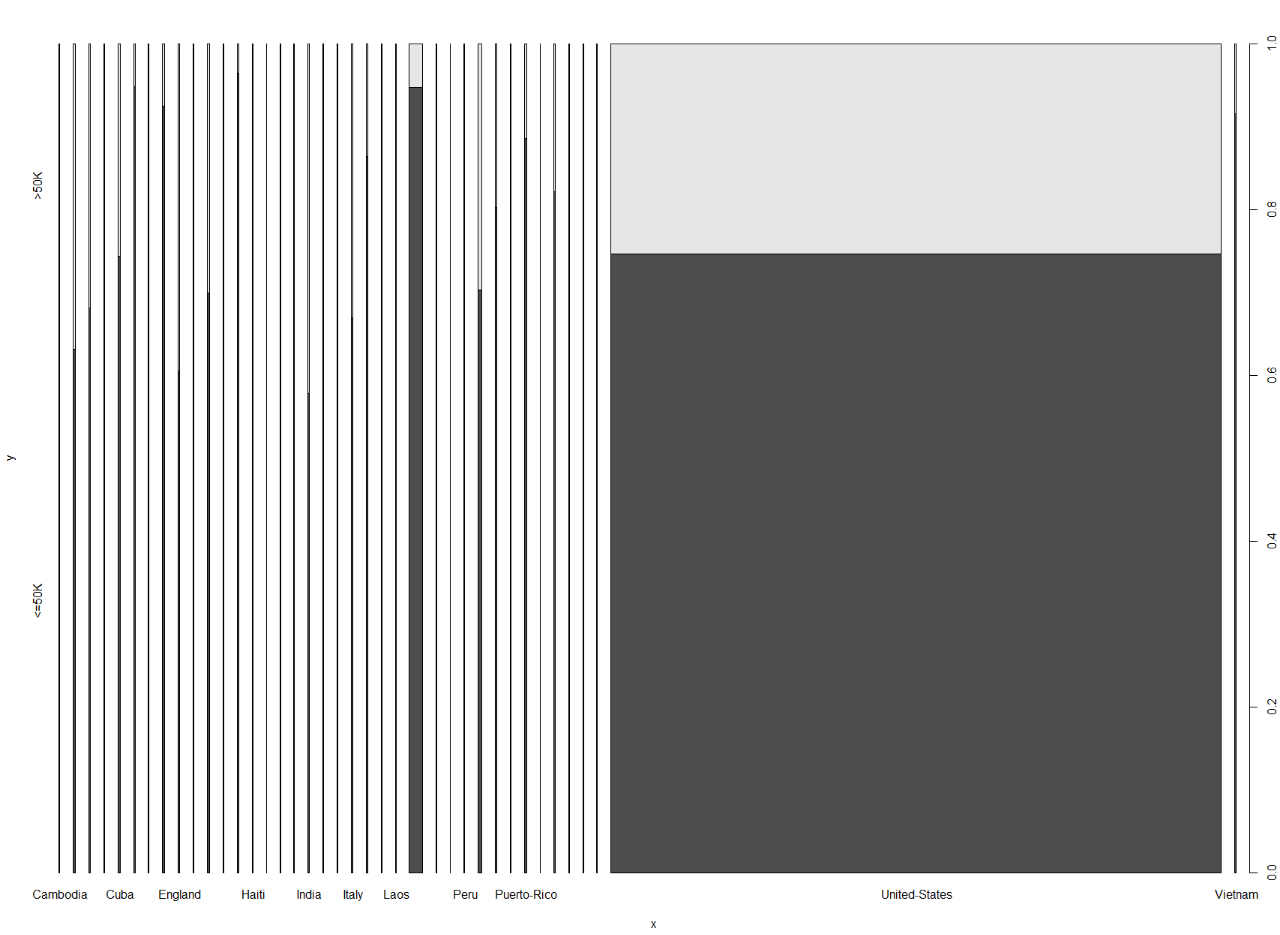
* Similar to capital gain it’s only 25% of people in each category who have beared any capital loss.

1. Plot of hours per week against salary:

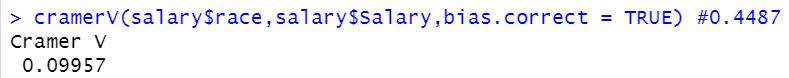


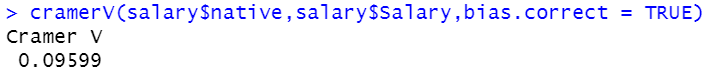
* We can infer from above plot that 75% of people from both the classes are working in between 35 to 50 hrs. /week.

1. Plot of native against salary:



* We can infer from above plot that most of the people belong to the United States.
* Feature selection:
* By using Cramer’s V and Uncertainty Coefficient I found out that the categories race and native are not much correlated with the output variable salary.





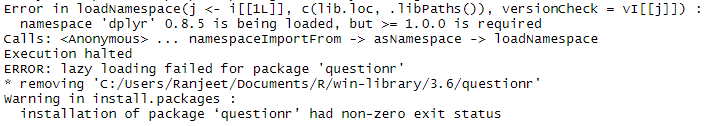




* So, I decided to remove those variables.
* Applied naive Bayes model and got and accuracy of 81.76%.
* After using a Laplace factor with a value of 15, I received an accuracy of 81.73%.
* We can keep optimizing the model by changing the Laplace value.
* We can use class imbalance and hyper-parameter tuning by using cross validation technique. Unable to use the algorithm:

“model = train(train\_x,train\_y,'nb',trControl = trainControl(method = 'cv',number = 10))”

Due to package installation issue. This algorithm runs using the ‘klaR’ package and for that it need ‘questionr’ package. I did try to find on Google the solution for this, but was unable to find. Error I’m facing:



* But I tried doing it the other way by using the algorithm:

model <- naiveBayes(salary\_train\_new$Salary~.,data = salary\_train\_new,trControl = trainControl(method = 'cv',number = 10))

And I received an accuracy of 82.19%.

* We can keep improving the accuracy further by tuning the parameters, and by knowing the parameters to tune.