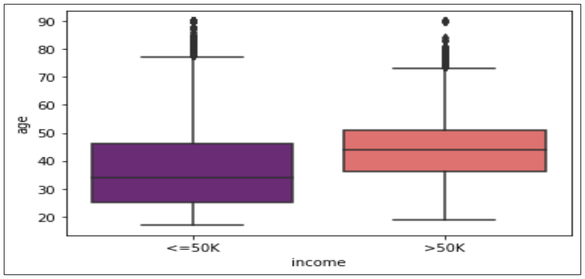
1. Visualizations for EDA

In the data given, it has the dependent variable as income and 14 independent variables such as age, workclass, education and so on. Here are some basic visualizations to gauge which of the variables can impact income. As expected, the age seems to have a positive relation with income and so does the hours per week, as shown in the below boxplots



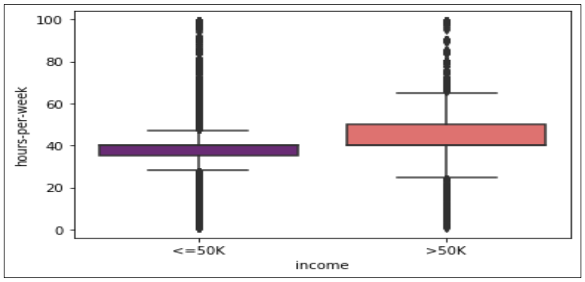
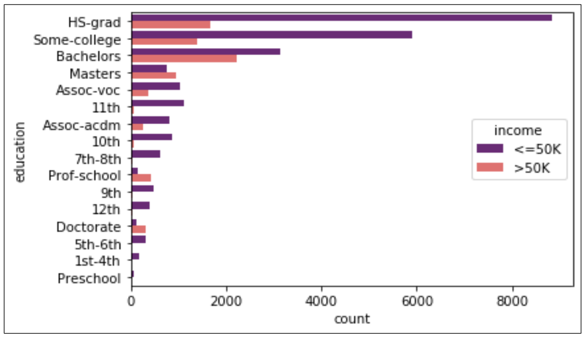


Fig 1.1 – Age v/s income

Fig 1.2 – Hours per week v/s income

Contrary to the description in the data dictionary, when plotted as factors education-num and education yield the same graph. Which means education-num does not represent the number of years of education but the ordinal representation of education type.



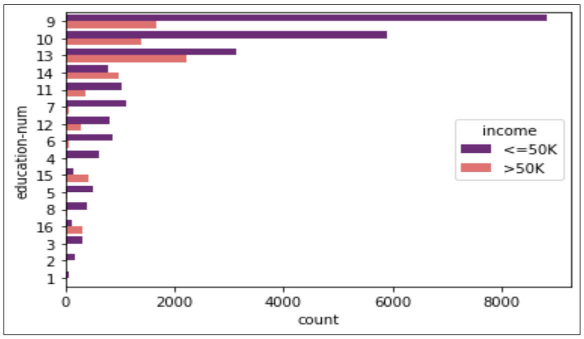
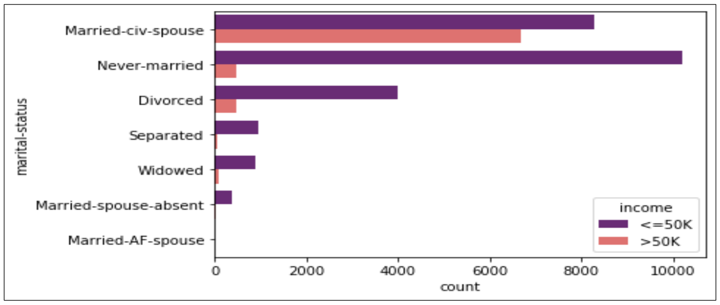


Fig 1.3 – education level countplot

Fig 1.4 – education-num countplot

There is a clear indication that marriage status (Fig 1.5) affects income. We see that women tend to earn less in numbers and the fraction of women who earn high is lower to that of men (Fig 1.6). Another determinant of income is job, we see levels in workclass(Fig 1.8) & occupation (1.7) are good predictors.



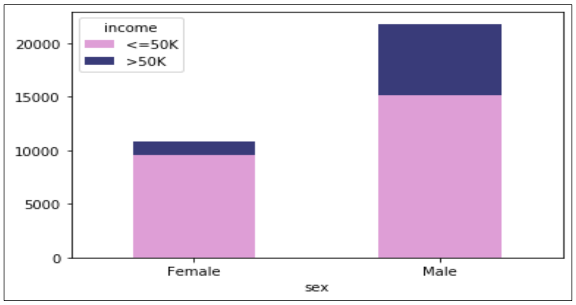


Fig 1.5 – Marital Status countplot

Fig 1.6 – Gender v/s income

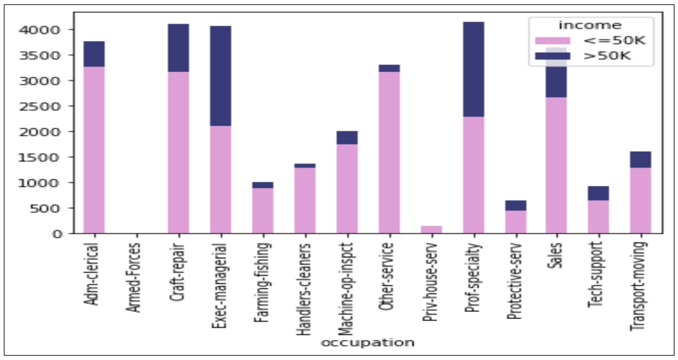


Fig 1.7 – Occupation stacked plot

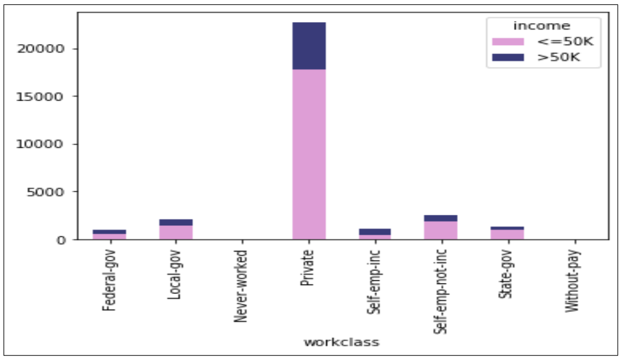


Fig 1.8 – work class stacked plot

Visualizing (Fig 1.9) capital gain and capital loss as it appears from the data that they are complementary as in no observation has both capital gain and capital loss. The yellow indicates non-zero values. In Fig 1.10, we looked at the missing data (yellow bars) and see that for workclass and occupation the data is not missing at random therefore we do imputation.

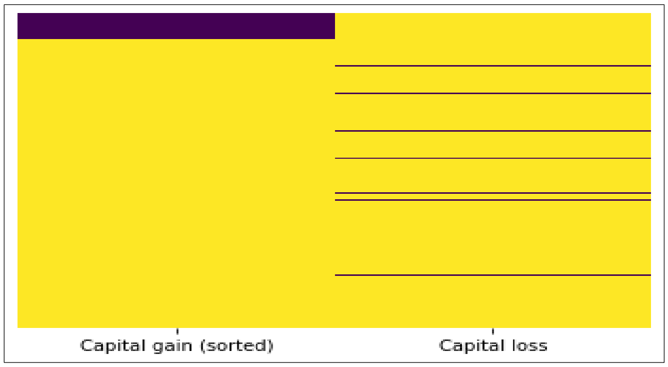


Fig 1.9 – Heatmap Capital gain and loss graphically

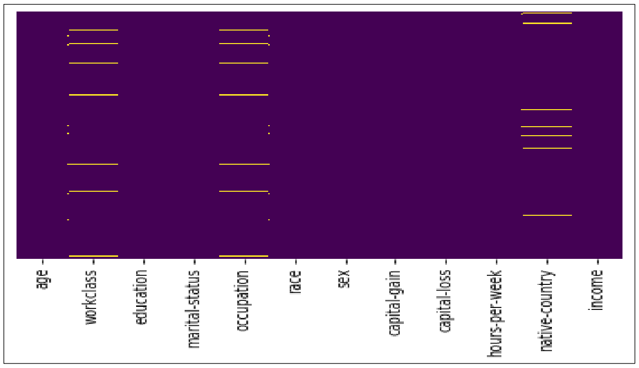


Fig 1.10 – Heatmap of missing data

2. Variable transformations

Post imputation in place of '?' in the data, we will transform the variables as per the table below

Variable name	Transformation	Reasoning
Age, Hours per week	Keeping as is	It is a continuous variable
Capital gain/loss	Capital Change is the difference	As preempted in fig 1.9, we combine the two by taking diff
Workclass	Level reduction to Self-employed, government, private and others	As indicated in fig 1.8, we see skewness towards private, to balance we combine the government and self-employed sectors.
fnlwgt	Dropping it	This variable captures the already known socio economic factors, we believe that it has no add on value as a predictor
education	Level reduction as per the given code	combine all the levels before high school as they stand the same from job readiness standpoint
Education-num	Dropping it	(fig 1.3 & fig 1.4) same data as education level
occupation	Keeping as is	We see that this variable is fairly distributed
relationship	Dropping it	High overlap with gender and marital status data

Race, gender	Keeping as is	We will keep the demographic data intact, also in case of race there has been aggregation in the given data.
Native-country	Levels reduction to US & non-US	This data is predominantly from US

Table 2.1 – Variable transformation prior to modeling

3. Feature elimination

Through EDA and visualization, we considered 9 out of 14 predictors in our model. With this first level of elimination and variable transformation, the AUC (Area Under the Curve) has reached 0.89 (Fig 3.1). AUC, a critical criterion to evaluate the model, can illustrate the ability of a binary classifier model regardless of the balance of data. Even though the AUC can to some extent handle imbalanced data, we still oversampled people whose income is more than 50k.

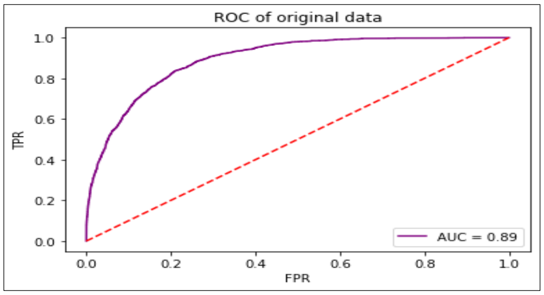


Fig 3.1 – ROC post EDA transformation

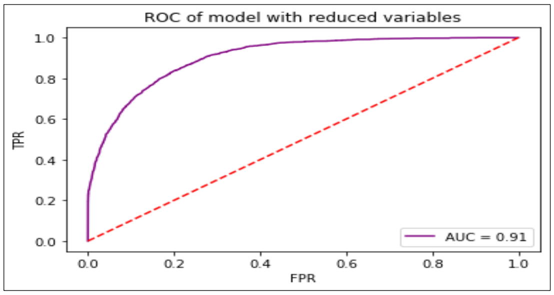


Fig 3.2 – – ROC post EDA transformation and feature selection

Then, based on the existing variables, we used the decision tree model (Fig 3.3) for feature selection. The tree only splits when an attribute can easily separate two classes. As a result, significant variables will be picked out by the tree, which in our case are marital status, age, education and capital change (net capital gain). With this change we see that that AUC improved to be 0.91 (Fig 3.2).

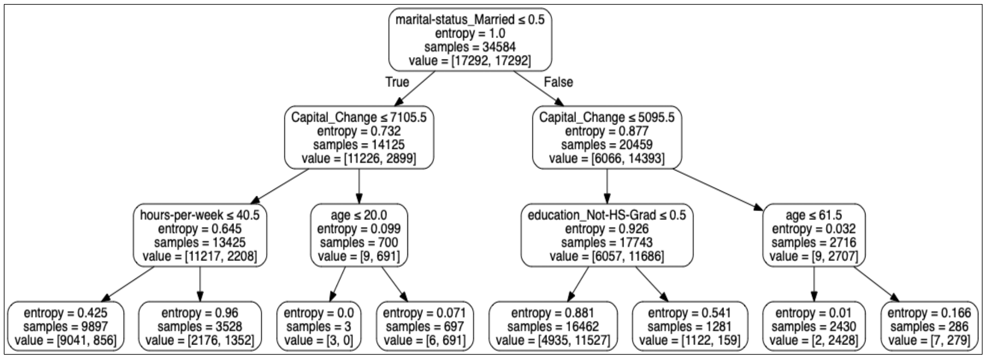


Fig 3.3 – Decision tree for feature selection

4. Regression model with income as dependent variable

Logistic regression model was used at first as this is a binary classification problem. We regressed income against nine parameters, and we corrected for oversampling. We eliminated the insignificant variables through the decision tree and ran another logistic regression model with four independent variables. This model fitted test data better with 83% f1 score, 85% precision and 82% recall while AUC improved from 0.89 to 0.91 (Fig 3.2). Instead of evaluating the model performance solely by its accuracy rate, we rely more on AUC score because the accuracy rate is biased with imbalanced date.

5. Regression model with work hours as dependent variable

In our first attempt, we built the model with all the variables and some interactions of interest. The adjusted R square for the base model is 17.2%. Later on, we deleted all the insignificant continuous variables (p-value ≥ 0.05) and merged insignificant levels based on their economic meaning, for example, splitting occupations into manual, intellectual and mixed occupations. However, we kept the categorical interactions that are partially significant, just discarded those that were completely insignificant. After modification, the adjusted R square dropped 14.3% due to fewer variables in the model, while all the variables excluding interactions were significant.

Parameter	Full model with insignificant variables	Model with only significant variables
Independent – variables	Age, workclass, education, marital-status, occupation, race, sex, native, income, net capital gain	Age, workclass, education, marital-status, occupation, race, sex, native, income, net capital gain
Interaction-variables	Sex-education, sex-occupation, native-race, age-workclass	Sex-education(Fig 5.1), sex-occupation (Fig 5.2), age-workclass (Fig 5.3)
Adjusted R-squared	0.172	0.143
AIC	250003	251081

Table 5.1 – Comparision between full and reduced model

The interactions we added here were the interactions between sex and education, sex and occupation, native and race, age and work class. We chose these interactions because we expected gender, race and age discriminations across different education levels, occupations, work classes and native countries. As the significant interaction plots show below, for example, work hours of male are mostly longer than

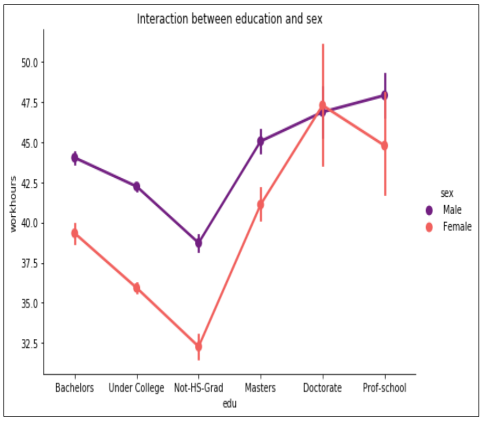


Fig 5.1 : Age and workhours

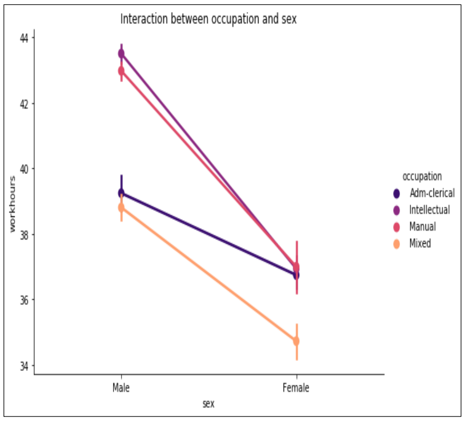


Fig 5.2: Gender and workhours

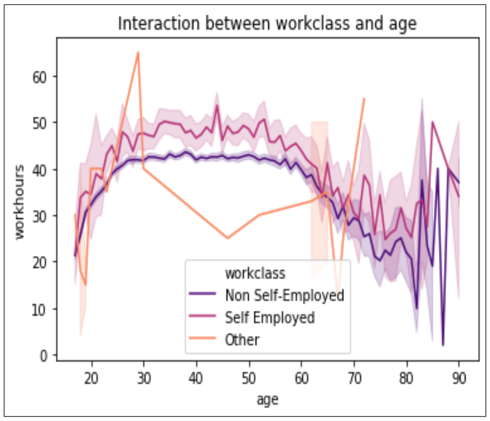
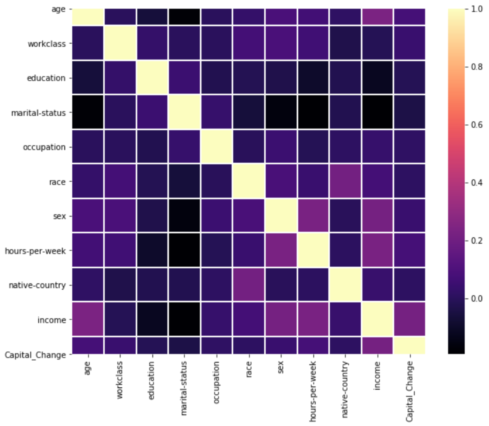


Fig 5.3 Workclass and age

6. Regression model with work hours as dependent variable for only sales occupation

Before running the regression, we first examined if there existed multicollinearity problem in the dataset. Heatmap implies that sex may be corelated with marital status to some extent, but it does not suffer from multicollinearity. Then, we included all variables and several interactions of our interest to



run the regression. It turned out that race is insignificant both by itself and with interactions and it was removed with other insignificant interactions. Similarly, we merge insignificant factors and create a new one with accordance to economic meaning. As a result, the Adj. R-squared remained to be 0.265 while AIC improves a little with all variables significant.

Compared to the Adj. R squared in Q5, the Adj. R square is much higher by only taking salesperson into consideration. It may be due to the fact that taking a subset of occupation lower the variance of the model, which improves the performance. Meanwhile, it is also possible that work hours of salesperson rely on these factors above, but there are unknown variables that greatly determine the work hours for people of other occupations

Parameter	Full model with insignificant variables	Model with only significant variables
Independent – variables	Age, workclass, education, marital-status, race, sex, native, income, net capital gain	Age, workclass, education, marital-status, sex, native, income, net capital gain
Interaction-variables	Sex-education, sex-race, native-workclass, age-workclass, income-age, income-sex, income-native	Education-sex (Fig 6.1), native-workclass (Fig 6.2), workclass-age (Fig 6.3), age-income (Fig 6.4)
Adjusted R-squared	0.265	0.265
AIC	28133	28111

Table 6.1 – Comparision between full and reduced model

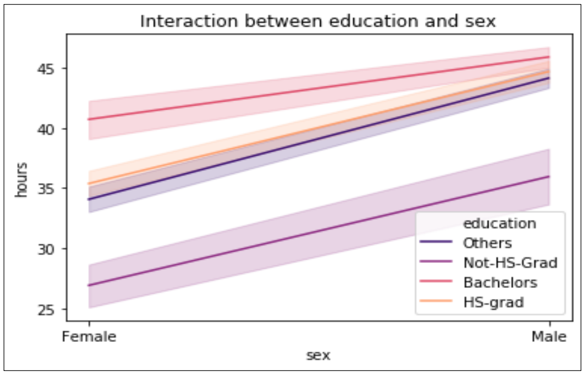


Fig 6.1 : Education and sex

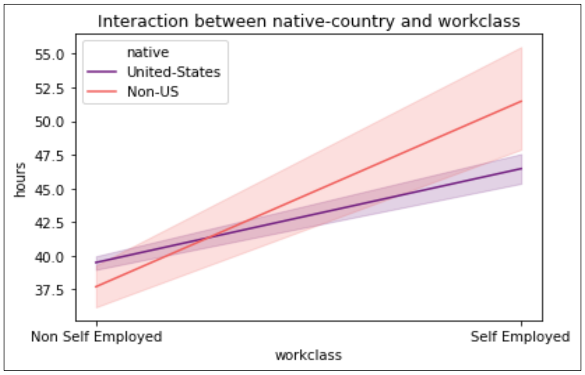


Fig 6.2 : Native-country and workclass

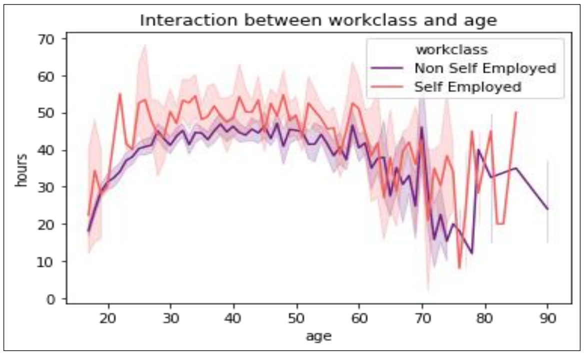


Fig 6.3 : Workclass and age

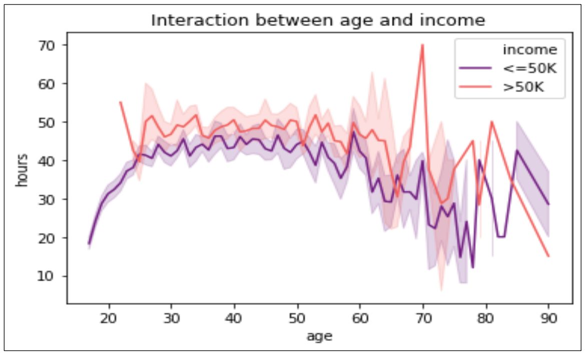


Fig 6.4 : Age and income