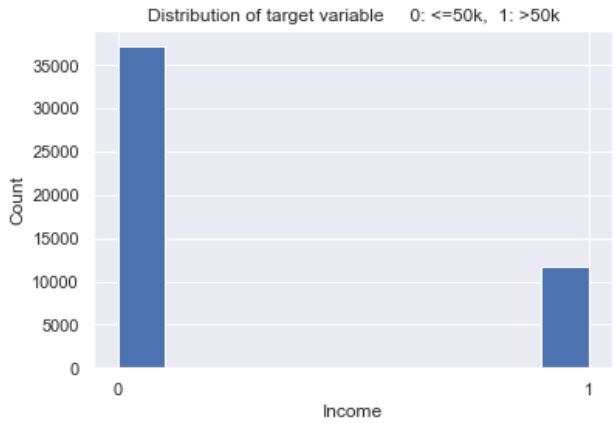
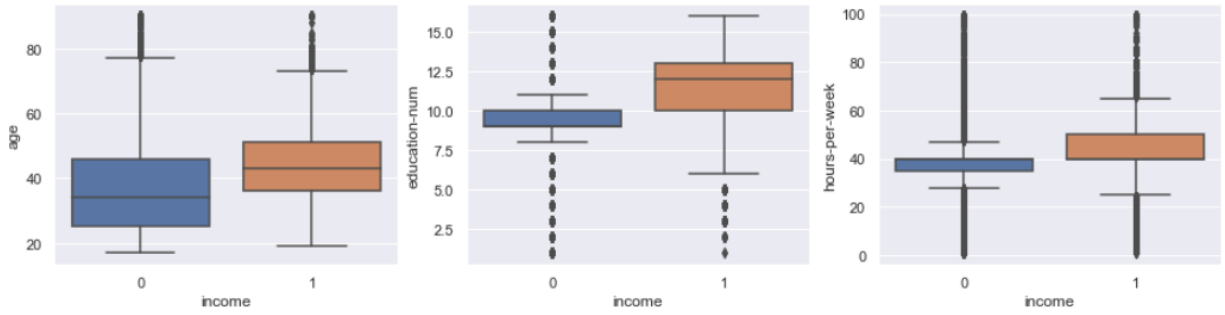
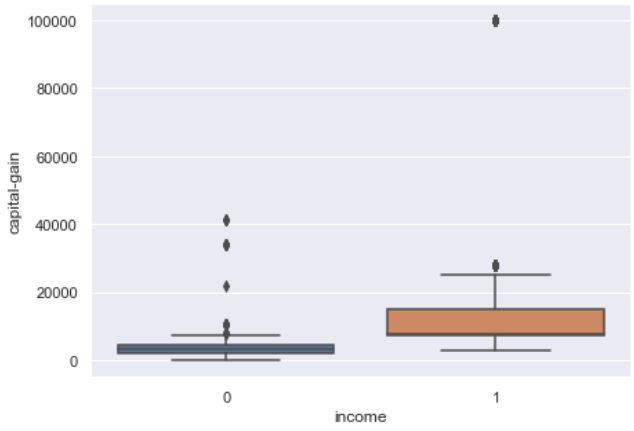
Capstone 2
Exploratory Data Analysis

We loaded the dataset from our previous data wrangling exercise. The distribution of our target variable "income" is shown below. About 75% are people with income less or equal than $50k, the remaining 25% have income greater than $50k

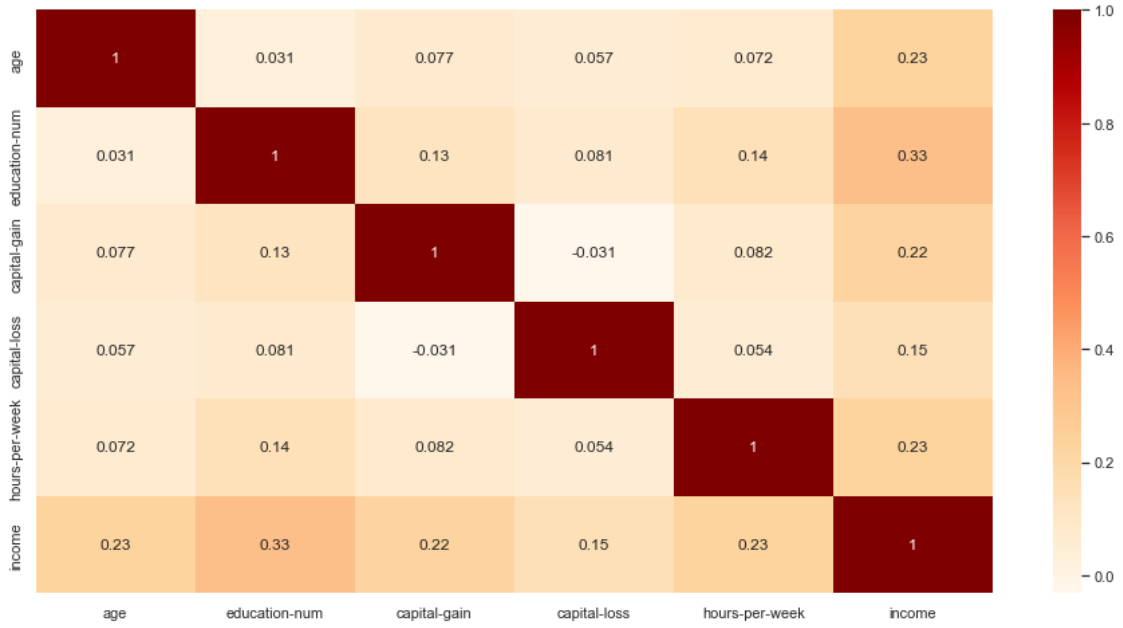


We then proceeded to do visual inspection of the numerical features. Education years, hours per week and capital-gain showed some differentiation against our income target variable. Also for age, but less pronunciated. Capital loss didn't show much variability.



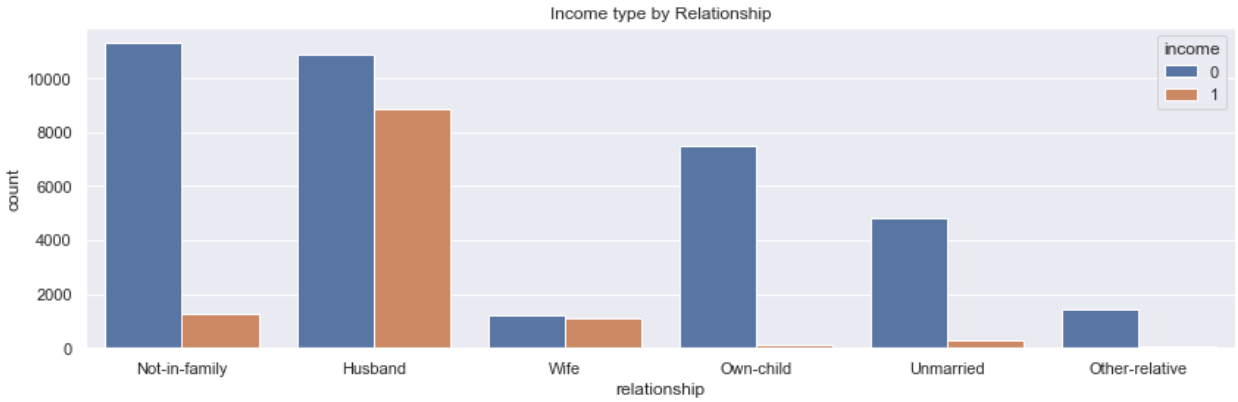


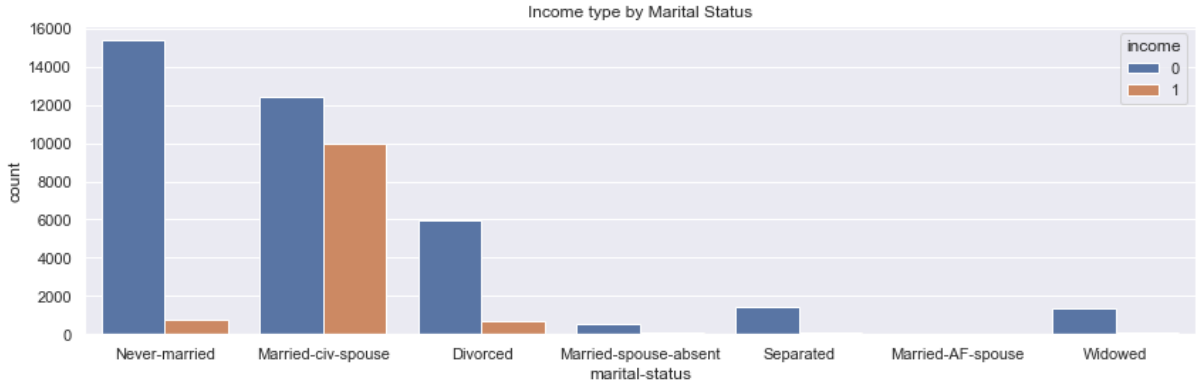
Correlation heatmap for numerical variables also confirmed these 4 features having some correlation with target variable income.

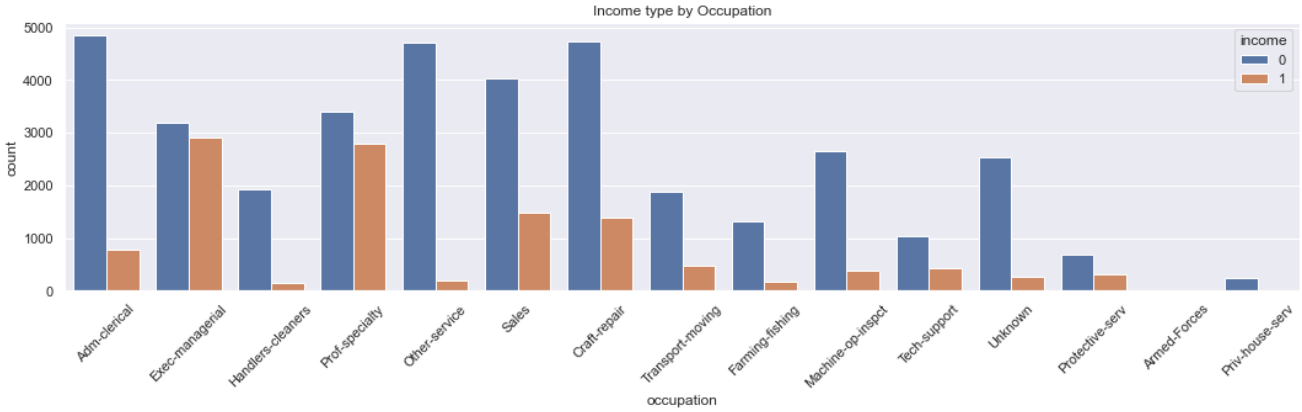


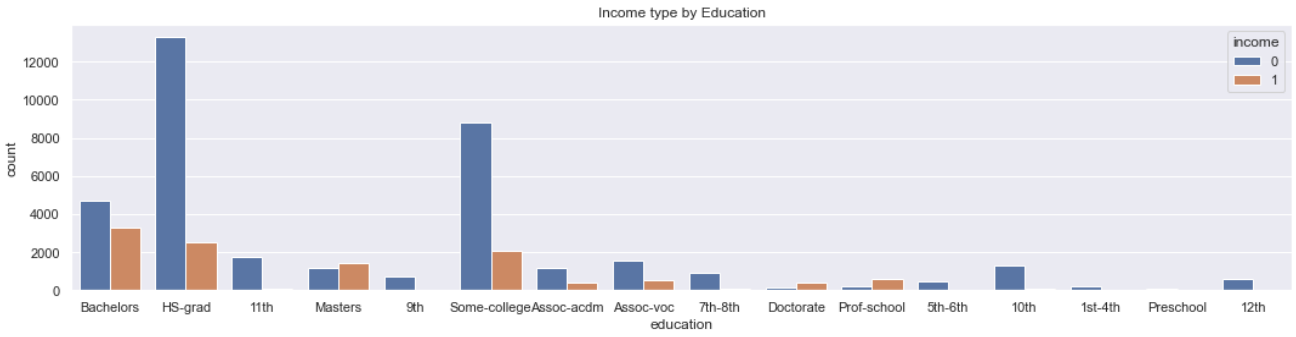
Before proceeding to visually inspect our categorical features, we added 2 more features: age range and work schedule based on hours worked per week.

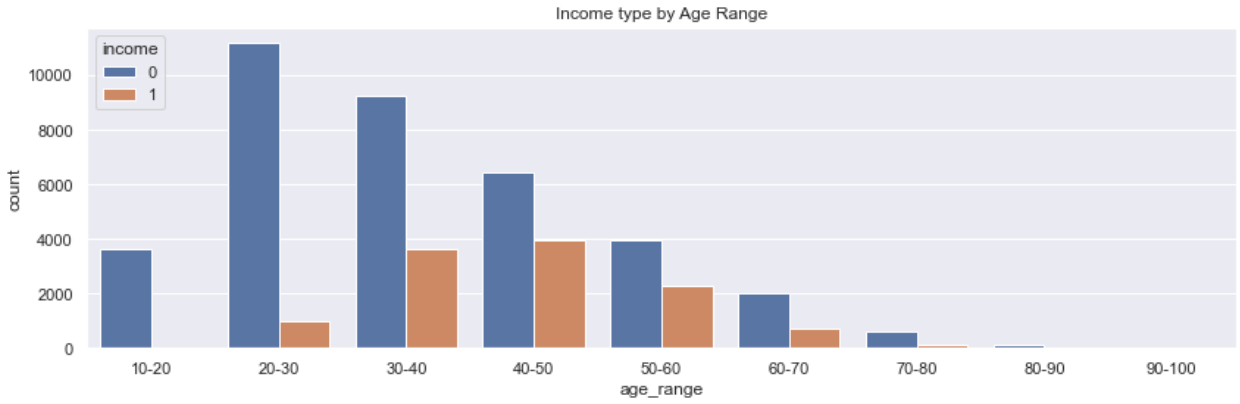
We then proceeded with visual charts for our categorical variables with the following features being the most relevant: Relationship, marital status, education, occupation and age range.



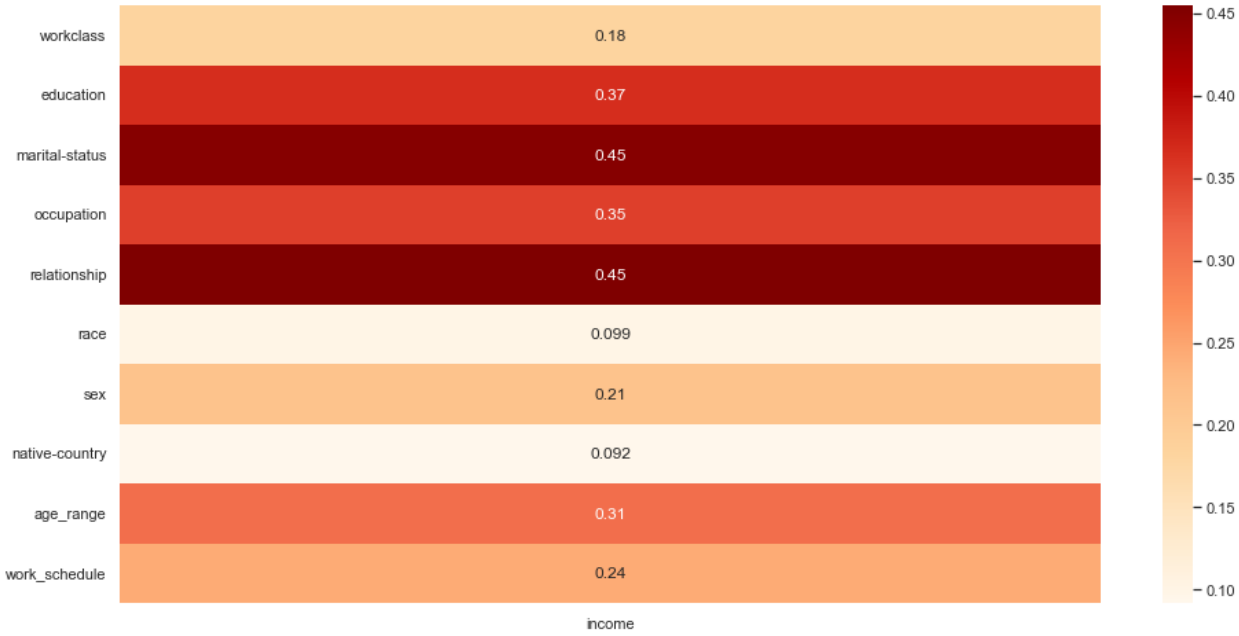

Income type by Marital Status
income
0
1
count
16000
14000
12000
10000
8000
6000
4000
2000
0
Never-married
Married-civ-spouse
Divorced
Married-spouse-absent
Separated
Married-AF-spouse
Widowed
marital-status


Income type by Occupation
income
0
1
count
5000
4000
3000
2000
1000
0
Adm-clerical
Exec-managerial
Handlers-cleaners
Prof-specialty
Other-service
Sales
Craft-repair
Transport-moving
Farming-fishing
Machine-op-inspct
Tech-support
Unknown
Protective-serv
Armed-Forces
Priv-house-serv
occupation

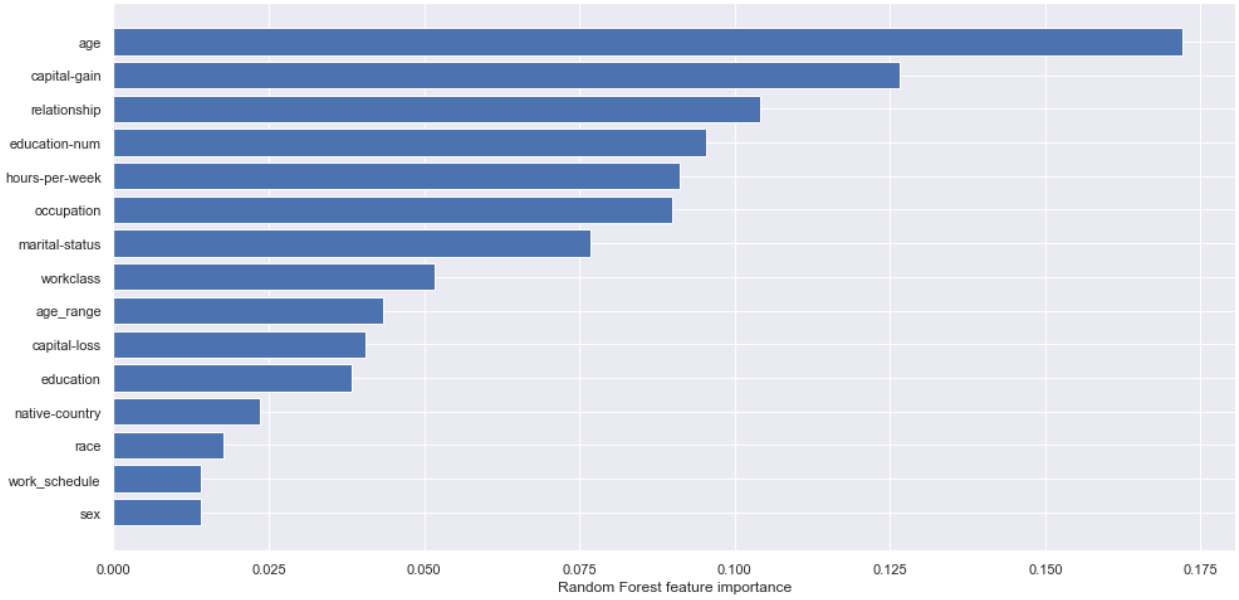

Income type by Education
income
0
1
count
12000
10000
8000
6000
4000
2000
0
Bachelors
HS-grad
11th
Masters
9th
Some-college
Assoc-acdm
Assoc-voc
7th-8th
Doctorate
Prof-school
5th-6th
10th
1st-4th
Preschool
12th
education


Income type by Age Range
income
0
1
count
10000
8000
6000
4000
2000
0
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
age_range

Correlation heatmap for categorical variables confirmed these features having some correlation with target variable income.



Finally we built a basic Random Forest model to obtain the features importances, with the following results:



This gave us a different point of view of some features we had seen before, but not strongly correlated with our target variable income. Age and capital gain had some correlation, but we can see them here as the top 2 features with more importance.
The 2 features we added (*workclass & age range*) might be redundant, since their original features have more importance than the bucketed new features.