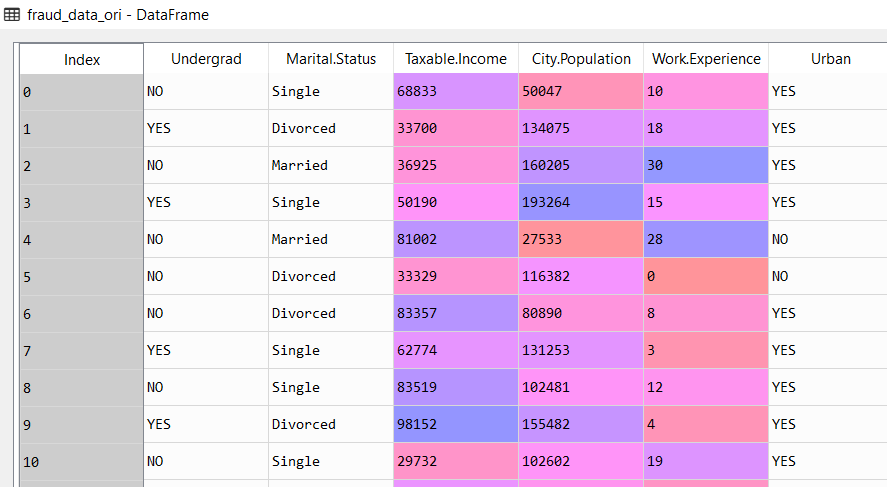
# Fraud check data in Random Forest Assignement

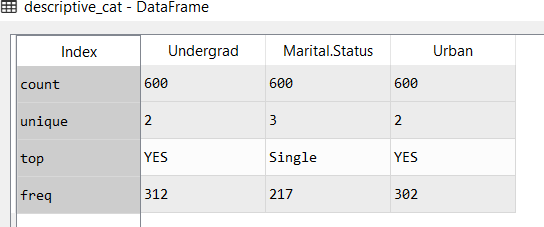
Use Random Forest to prepare a model on fraud data

treating those who have taxable\_income <= 30000 as "Risky" and others are "Good"

The following is the dataset:

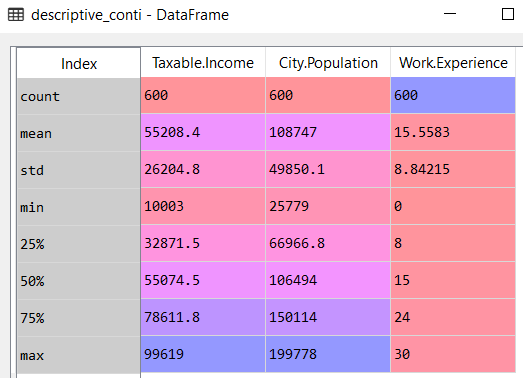


The descriptive statistics of categorical data is as follows:



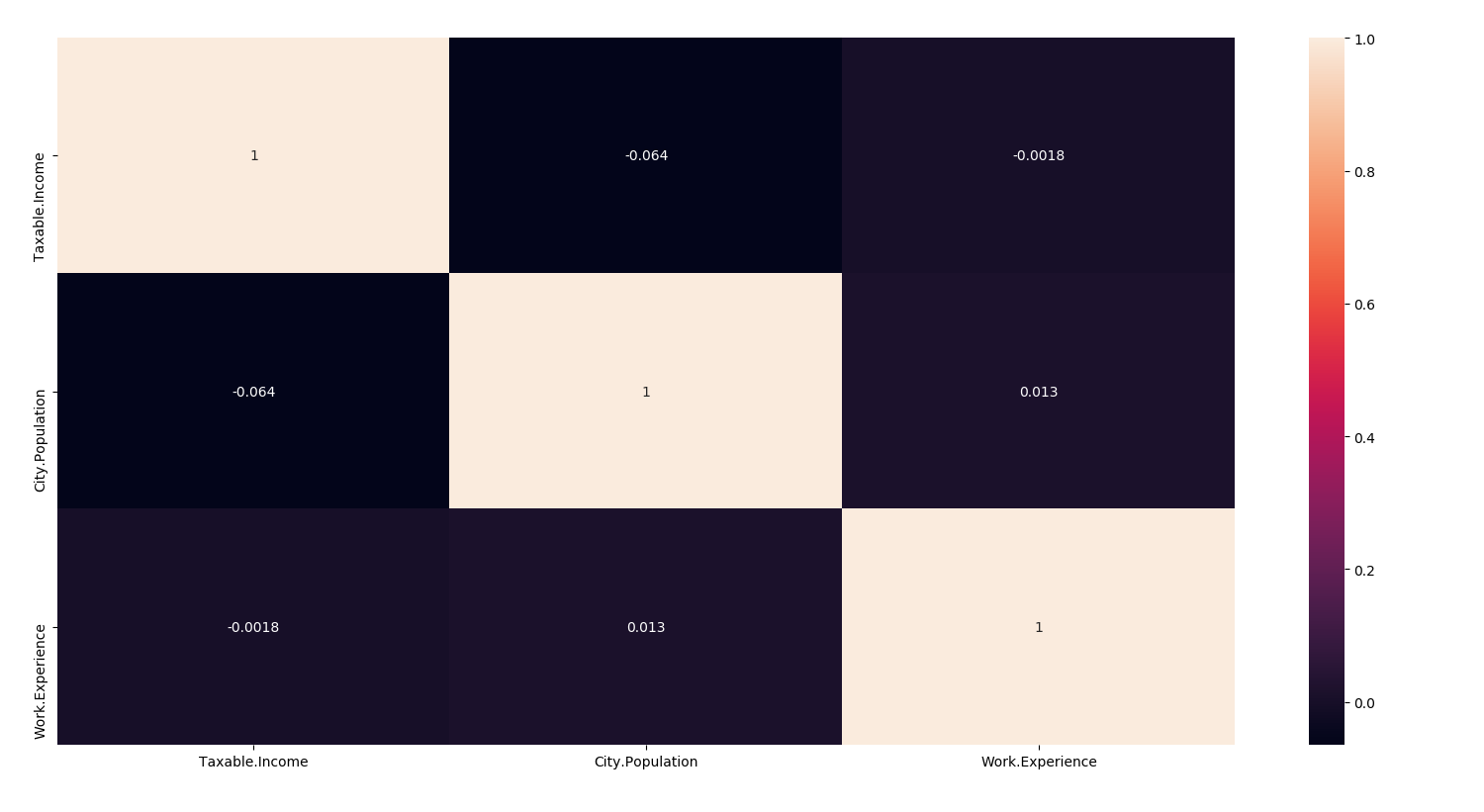
From the above data most of the individuals are single, they live in an Urban area and have an undergraduate degree

The following is the descriptive statistics of the continuous vairables



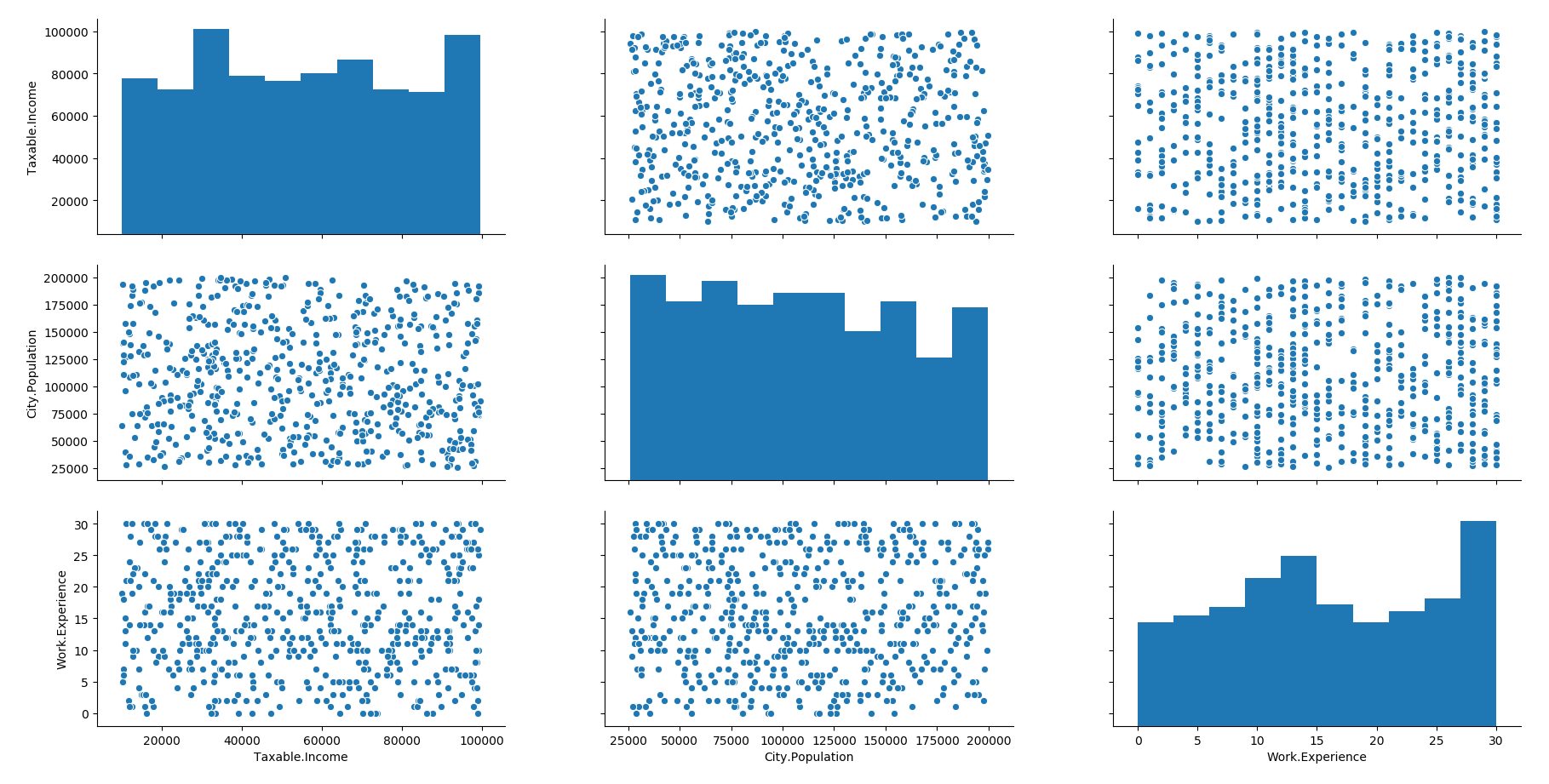
In the above details, the average work experience is 15 years, also the median work experience is also 15. More than 50% of the individual are above 50k Taxable income which means that more than half of the people fall under high risk category

The following is the heat map representing the categorical variables:

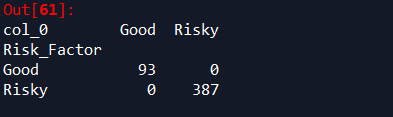


All the above variables have very low correlation. This mean that there is no co-variance issue in the dataset

The following is the pairplot of continuous variables



We feed the training data set of crime data into the random forest algorithm. The following are the results:



The accuracy obtained is 100%

The following are the results of the training data:

