Springboard Data Science

Capstone Project 1

Final Project Report

Predicting >50K and <=50K Income Class

Project Goals:

This capstone project is based on a dataset called 'Adult Data Set' from UCI's Machine Learning Repository.

This project aims to solve following problems:

1. Predict whether a person's income exceed $50K/yr based on the census data
2. Identify various segments which may exist in the population based on the data collected

Target Applications of this Data Analysis:

Census data is almost like a gold mine for business users. Primary reasons being, it is a true population data which captures the information about the total universe of data instead of a mere sample.

Such census data when combined with income classification based on 50K+ threshold would form a very valuable source of population demographics. Once we analyze this data and establish potential relationships between some of the key demographic parameters which influence the income category, this information can be very effectively used in making many economic, social, marketing, advertising, sales promotion and many other similar business decisions.

I envisage that this research project would be very useful to search marketers, social media advertisers, automotive manufacturers (or similar big ticket products manufacturers and marketers), media and print advertisers, etc.

A little more information about the dataset:

Dataset Information: The dataset is based on the 1994 US Census Database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

Prediction task is to determine whether a person makes over 50K a year.

Attribute Information:

- income: >50K, <=50K.
- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education-num: continuous.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Project Plan / Approach to achieve the goals outlined above:

a. As in any good data analysis project, I first plan to do a good deal of EDA (Exploratory Data Analysis) of the dataset. The objective is to thoroughly understand the data and then based on that prepare the data analysis plan/framework.

b. Based on the EDA, I will identify the key hypothesis which should answer following questions

- Overall income distribution of the population
- Which attributes may have greater influence on determining the income threshold of 50K/yr
- Any other interesting facts

c. Use regression methods to determine which attributes have higher impact/influence

d. Plan how to use the visualization techniques learned to present the findings

e. Use Supervised learning methods to identify patterns, segments and attributes which drive the results

f. Use classification methods to classify income segments

g. Use ML techniques to predict the income level and compare the accuracy of the different models

h. Use Unsupervised learning methods to identify patterns, segments and attributes which drive the results

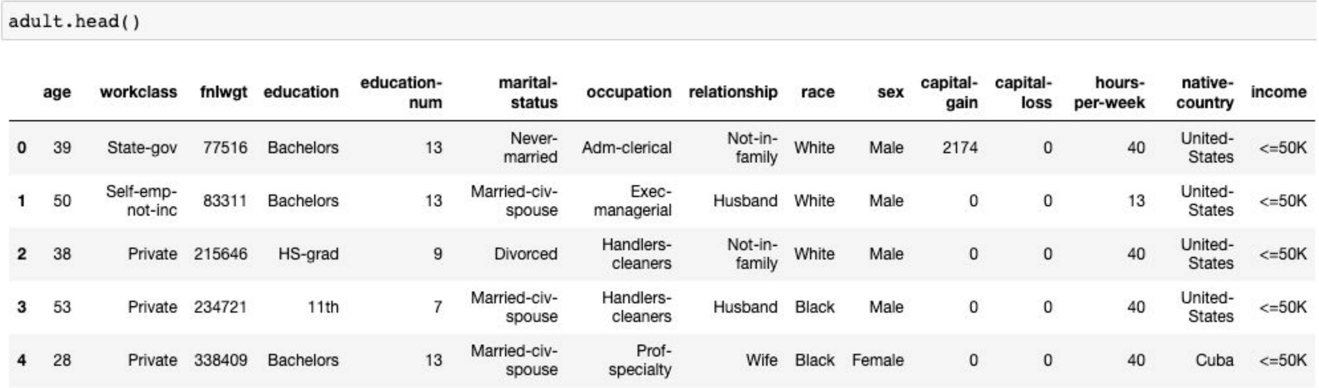
Key Deliverables, Collaterals

I envision following documents to document and present the result of this effort

- iPython Notebooks containing all the codes, analysis and results
- A summary/executive finding report
- Presentation slides

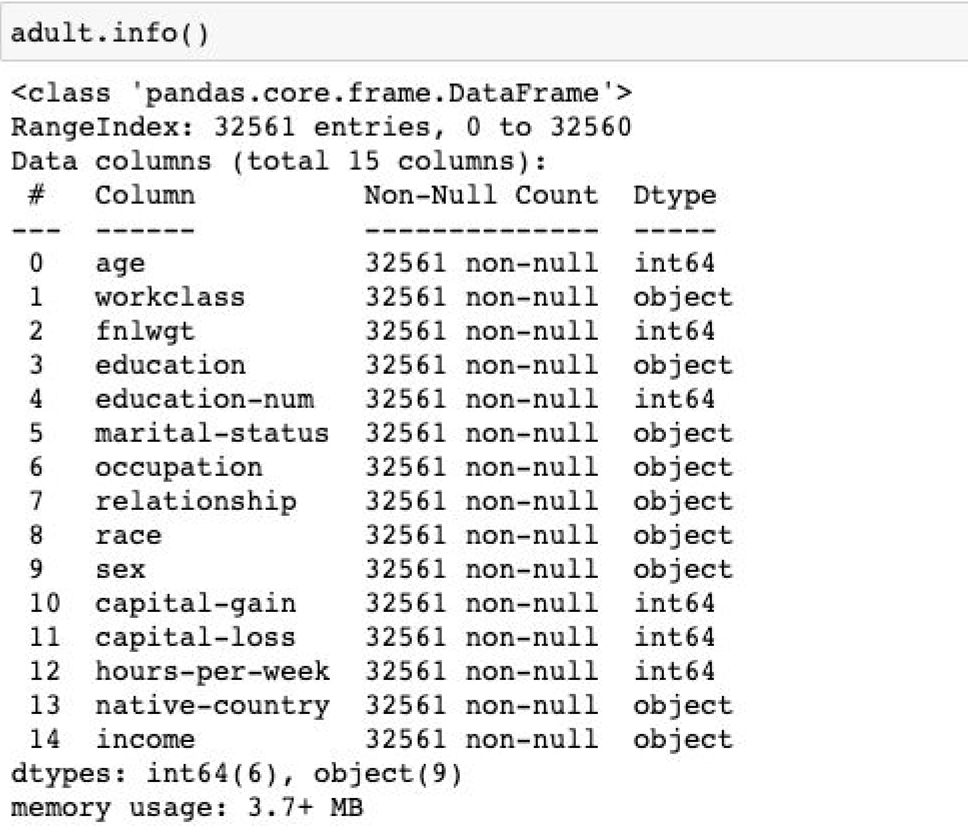
Data Wrangling

- The original data is in .data format. This was first converted to .txt, and then to .csv to make it easily readable by Pandas.
- The data has 32561 records (rows) and 15 attributes (columns).



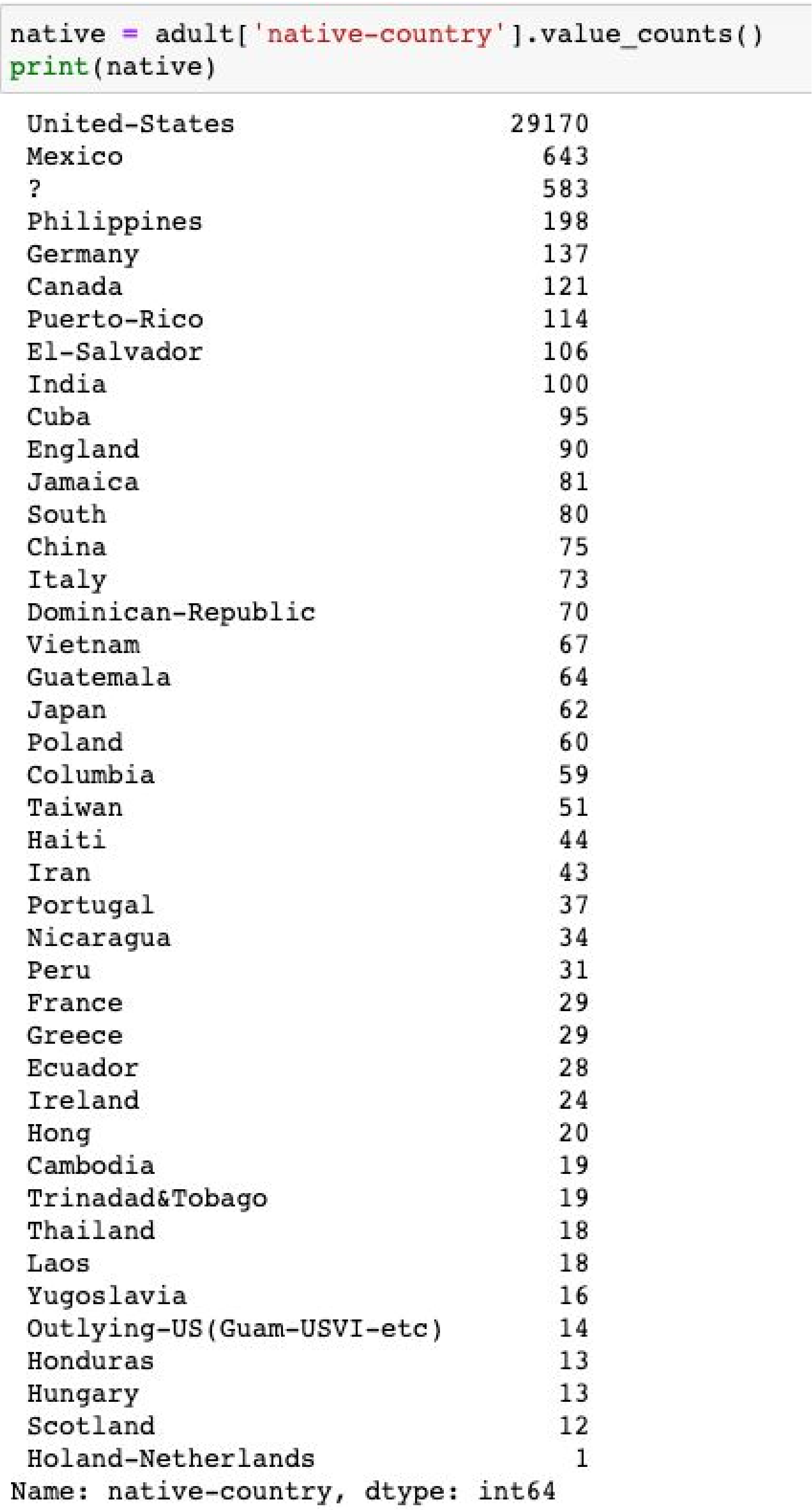
```
adult.head()
```

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K



```
adult.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   age             32561 non-null  int64 
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64 
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64 
 5   marital-status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital-gain    32561 non-null  int64 
 11  capital-loss    32561 non-null  int64 
 12  hours-per-week  32561 non-null  int64 
 13  native-country  32561 non-null  object
 14  income          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

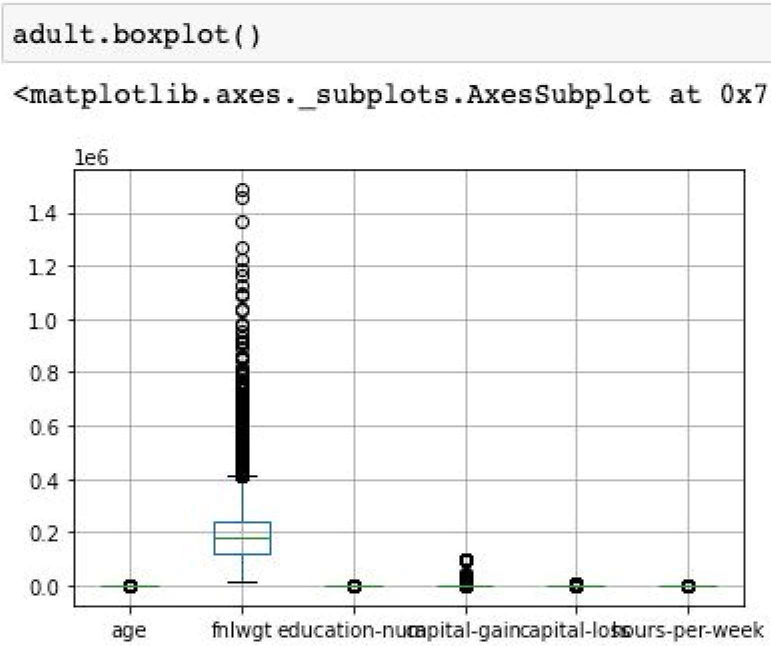


```
native = adult['native-country'].value_counts()
print(native)
```

```
 United-States                 29170
 Mexico                          643
 ?                               583
 Philippines                     198
 Germany                         137
 Canada                          121
 Puerto-Rico                     114
 El-Salvador                     106
 India                           100
 Cuba                             95
 England                          90
 Jamaica                          81
 South                            80
 China                            75
 Italy                            73
 Dominican-Republic               70
 Vietnam                          67
 Guatemala                        64
 Japan                            62
 Poland                           60
 Columbia                         59
 Taiwan                           51
 Haiti                            44
 Iran                             43
 Portugal                         37
 Nicaragua                        34
 Peru                             31
 France                           29
 Greece                           29
 Ecuador                          28
 Ireland                          24
 Hong                             20
 Cambodia                         19
 Trinadad&Tobago                  19
 Thailand                         18
 Laos                             18
 Yugoslavia                       16
 Outlying-US(Guam-USVI-etc)       14
 Honduras                         13
 Hungary                          13
 Scotland                         12
 Holand-Netherlands                1
Name: native-country, dtype: int64
```

- **Missing Values**?

There are no missing values, except native-country columns which have 582 entries with value '?' - for these entries, native-country was probably not known. Considering that the count is quite small compared to the total count (32561), I decided to leave it unchanged so that it can be identified as its own category in the later statistical analysis. If required, I can drop these rows later in the EDA and Data Story part of the project.



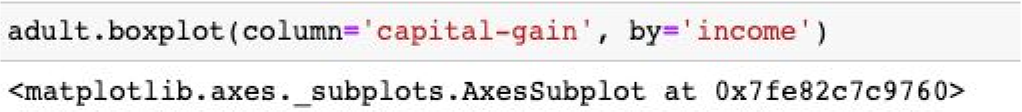
```
adult.boxplot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe82c5ff250>
```

The above box-plot cannot be used for EDA since the data is not normalized and as a result Box-plot cannot be displayed in a useful way. I will plot individual columns against each other to get more interpretable results of box-plots.

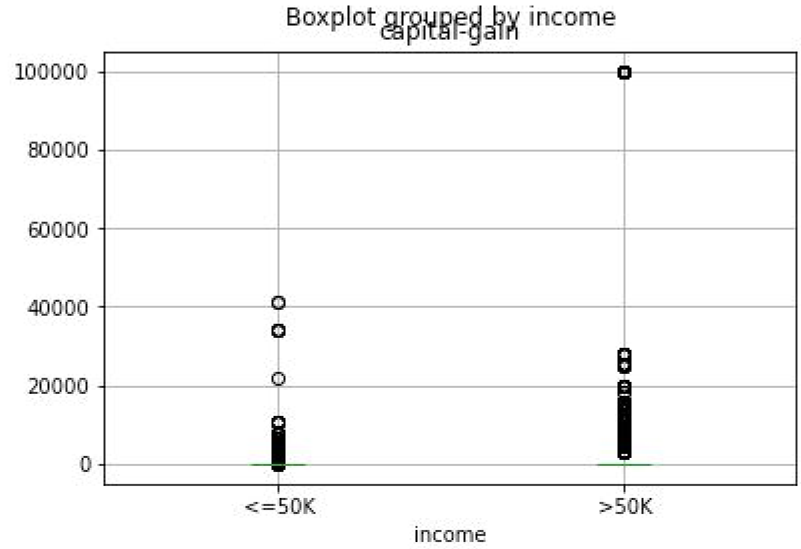
Also the column 'fnlwgt' values are not real numbers instead they are more of identification values (and hence of no use for quantitative data analysis). Hence in the later analysis steps, I dropped this column from the data frame.

Similarly, for Columns ‘Capital-Gain’ and ‘Capital Loss’ majority of the records had 0 value, hence in order to avoid any undue impact of these attributes on the overall data analysis and its results, I decided to drop these 2 columns as well from the table.



```
adult.boxplot(column='capital-gain', by='income')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe82c7c9760>
```



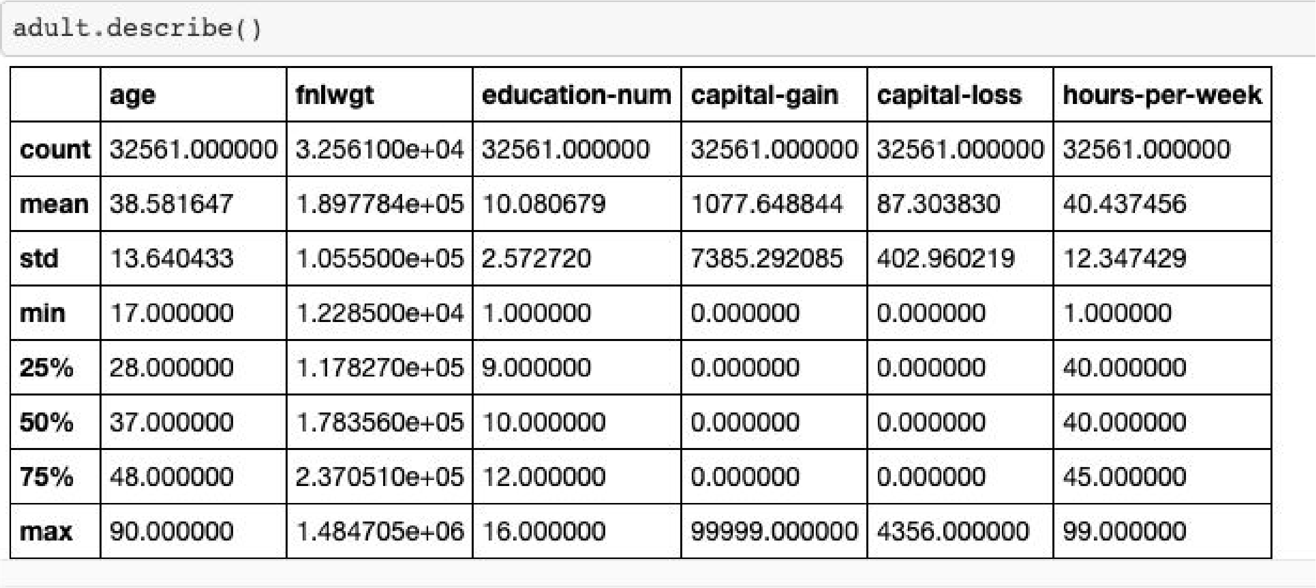
- **Outliers?**

The capital gain figure of $100000 may appear to be an outlier but it is quite possible to have that much income for some individuals. Similarly, there are a few outliers in the <=50k income group, but they also fall into the category of valid data/values.

As they say not all outliers are bad data points. Some can be an error, but others are valid values.

Data Story

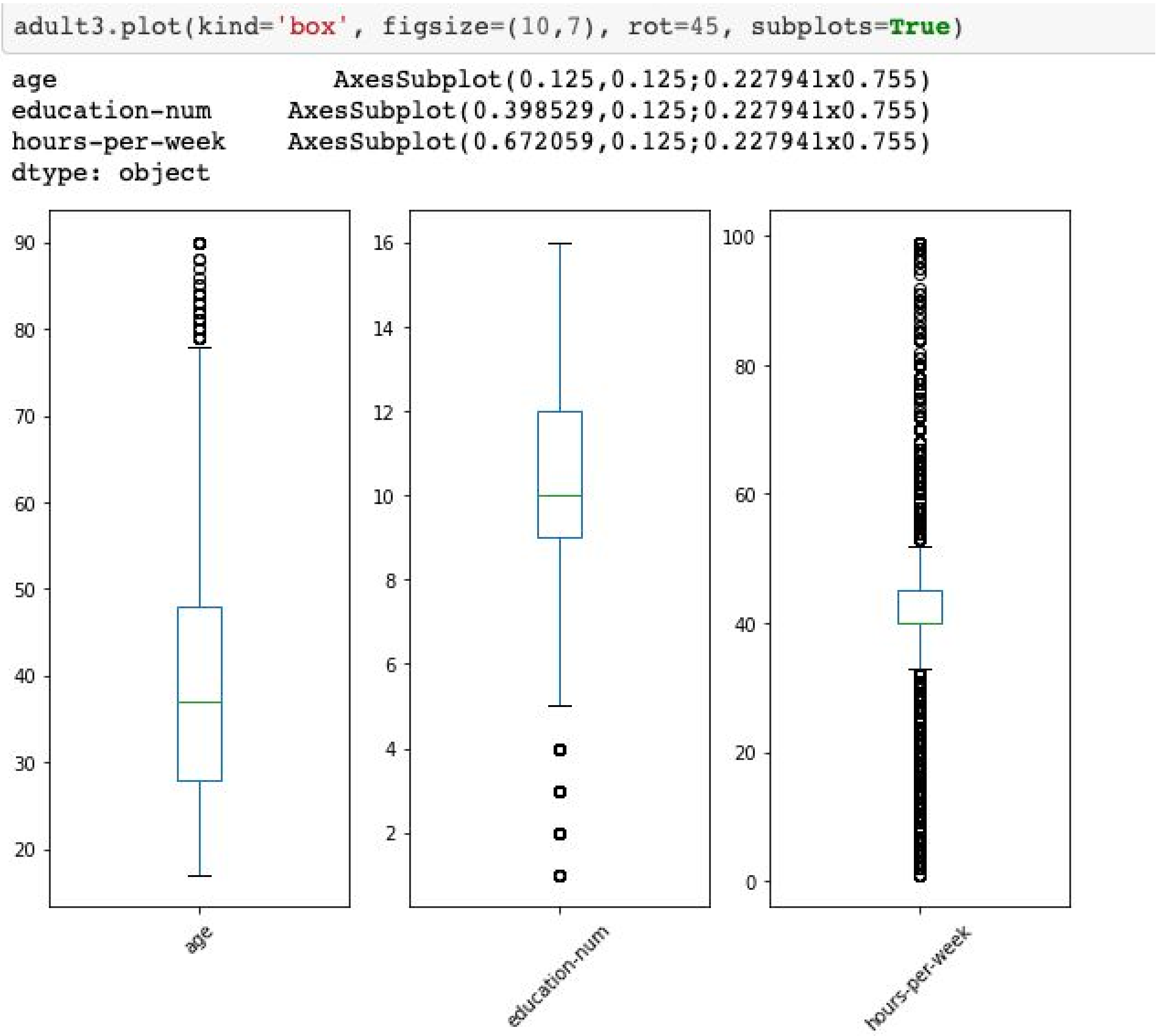
- First I tried to understand the statistics of the data using .describe() method of the data frame.



```
adult.describe()
```

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
count	32561.000000	3.256100e+04	32561.000000	32561.000000	32561.000000	32561.000000
mean	38.581647	1.897784e+05	10.080679	1077.648844	87.303830	40.437456
std	13.640433	1.055500e+05	2.572720	7385.292085	402.960219	12.347429
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.178270e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783560e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.370510e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

- I also used the box-plot method to visually analyze the distribution for integer type data attributes

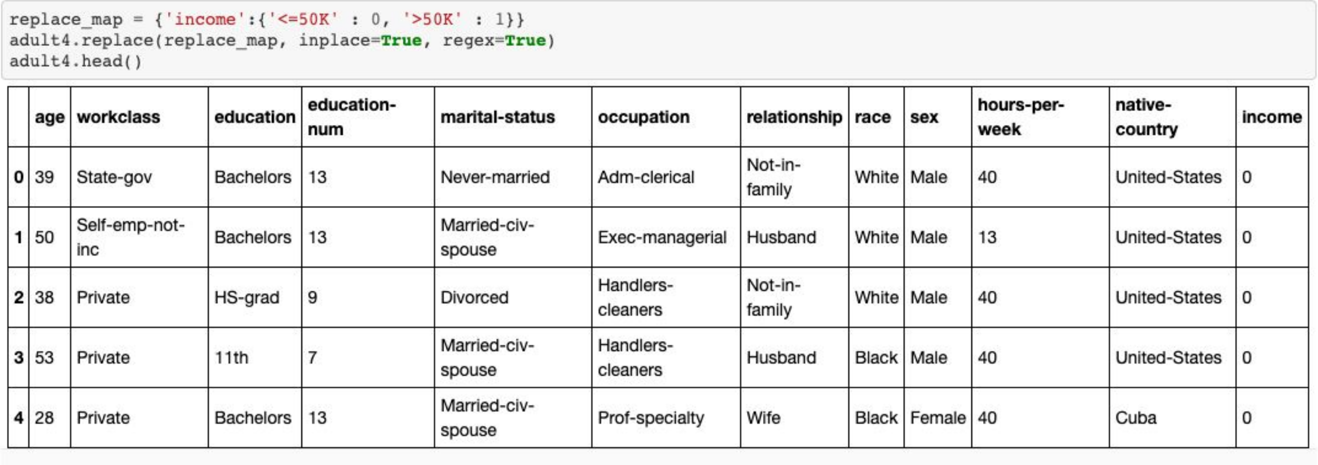


```
adult3.plot(kind='box', figsize=(10,7), rot=45, subplots=True)
```

```
age                AxesSubplot(0.125,0.125;0.227941x0.755)
education-num    AxesSubplot(0.398529,0.125;0.227941x0.755)
hours-per-week   AxesSubplot(0.672059,0.125;0.227941x0.755)
dtype: object
```

As part of further data wrangling and tidying steps, I also converted the non-integer attributes into ‘category’ type for future ease of processing and reduced processing time.

The Income attribute values which are either <=50K and >50K, were converted (transformed) to 0 and 1 (and integer type), so that this attribute can be used for further statistical analysis.

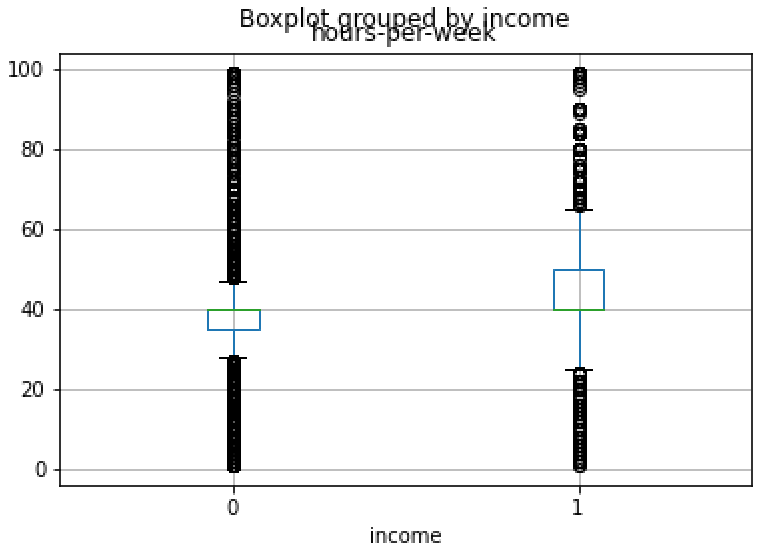


```
replace_map = {'income':{'<=50K' : 0, '>50K' : 1}}
adult4.replace(replace_map, inplace=True, regex=True)
adult4.head()
```

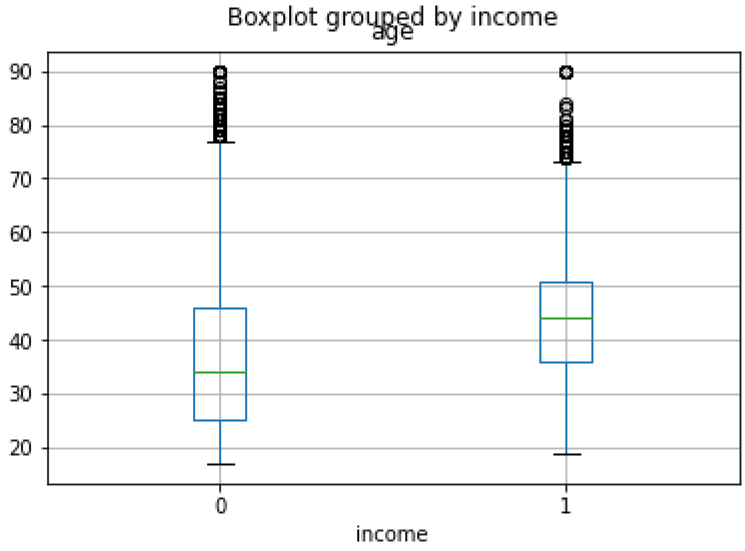
	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	hours-per-week	native-country	income
0	39	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	40	United-States	0
1	50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	13	United-States	0
2	38	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	40	United-States	0
3	53	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	40	United-States	0
4	28	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	40	Cuba	0

- **Following Are the Key Findings Based on Exploratory Data Analysis:**

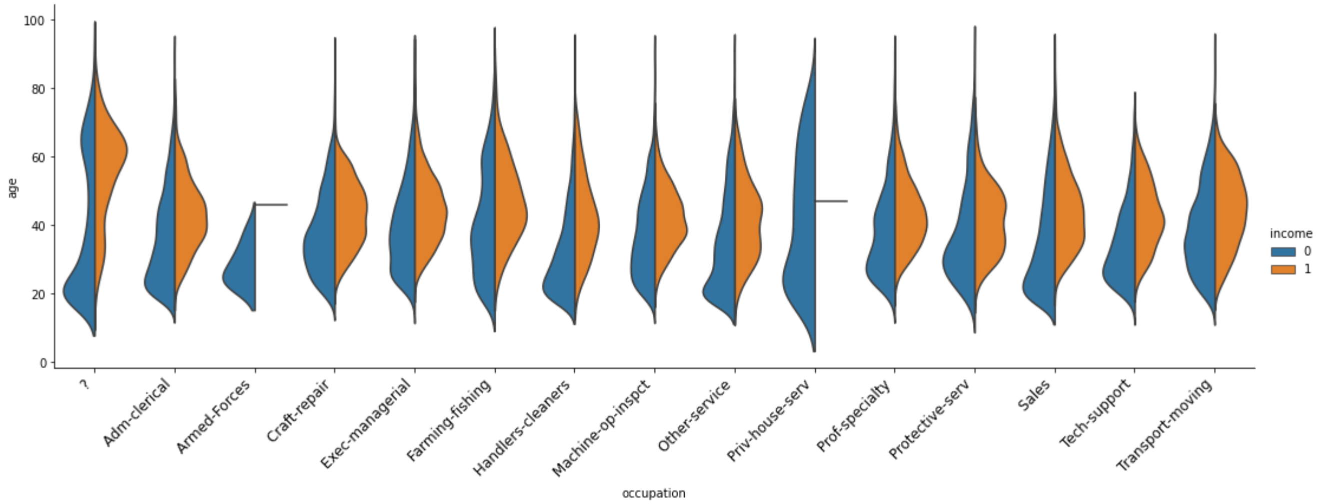
1. The low income group and high income group clearly differ from each other based on the data for attributes age and hours-per-week.
    a. The >50K income group clearly has people who are clocking many more hours than people from <=50K income group. The average # of hours are more than 40 hours in high income group vs. an average of less than 40 hours in low income group



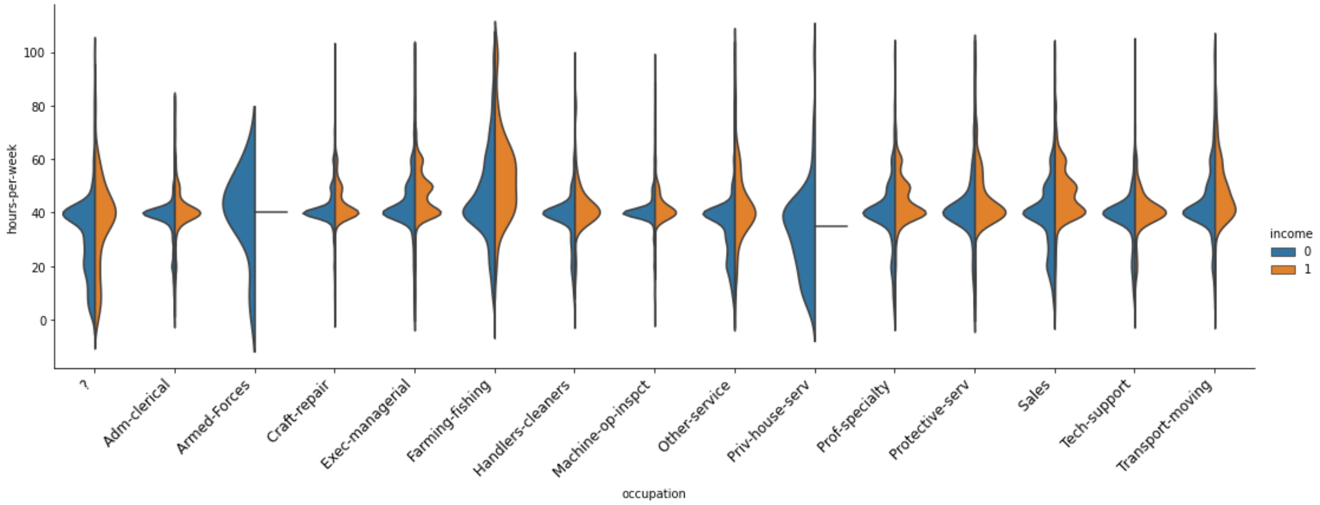
    b. The >50K income group clearly has people with higher age than people from <=50K income group. The average age is 44.25 years in high income group vis-a-vis average age of 36.87 years in low income group



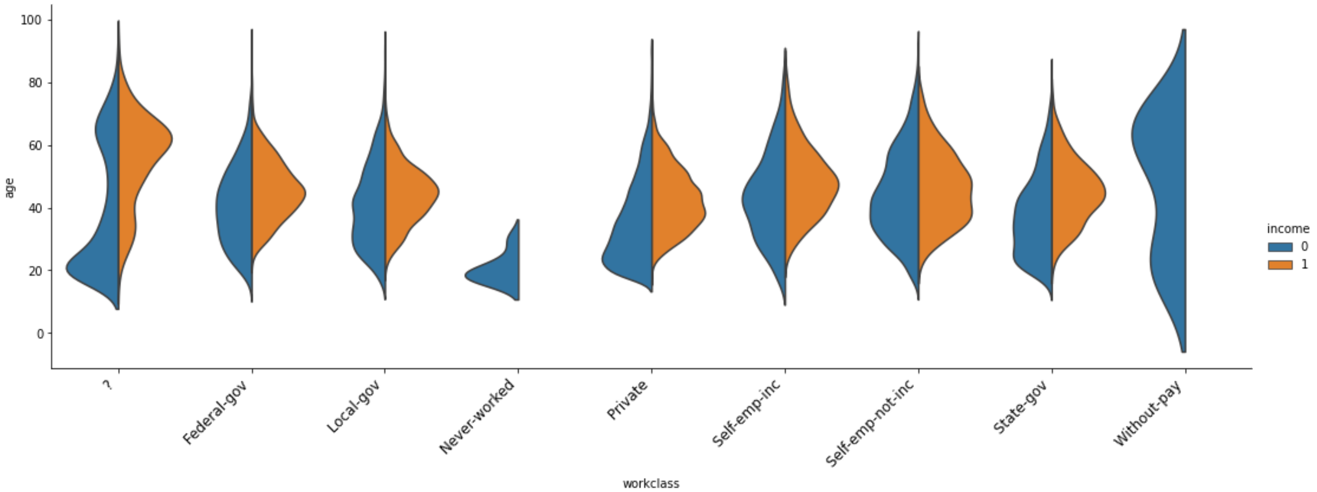
2. The overall trend of age difference between higher income group and lower income group continues across various occupation categories as well



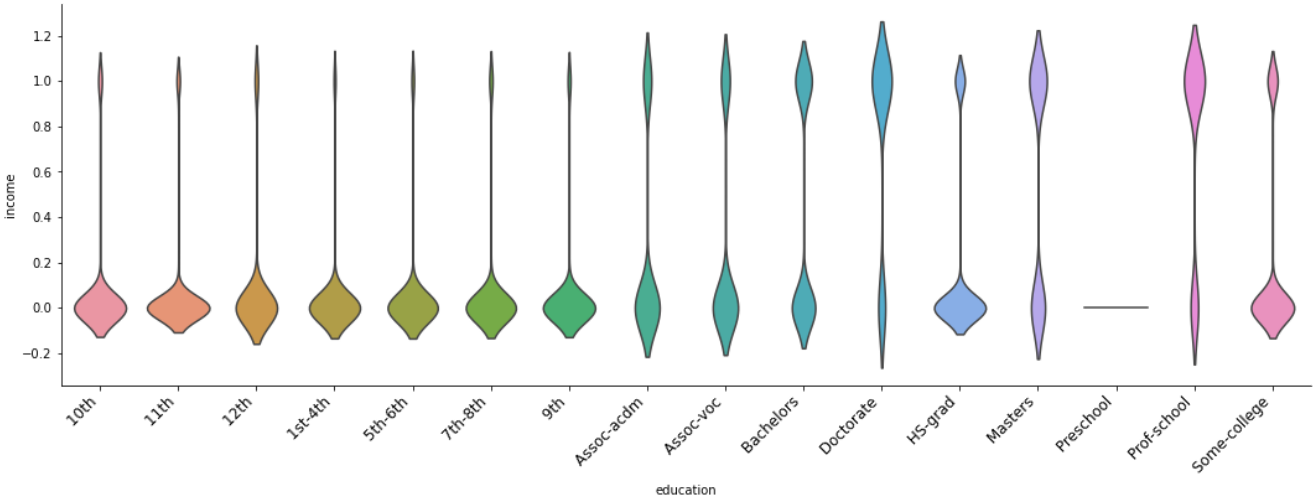
3. The overall trend of 'hours worked per week' difference between higher income group and lower income group continues across various occupation categories as well



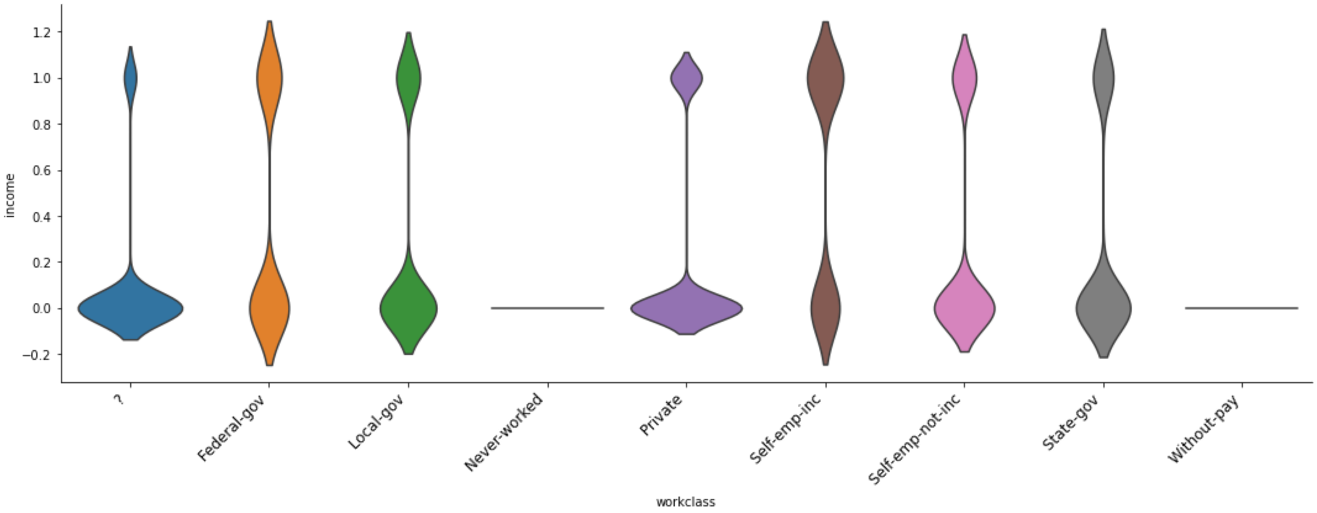
4. Across the work classes, for the higher income group, the age is more skewed towards higher values. The overall age difference between the low income group and high income group seems to be substantial. Lower income group is more concentrated between 20 to 40 years age block vs. higher income group which is more concentrated between 40 to 60 age block¶



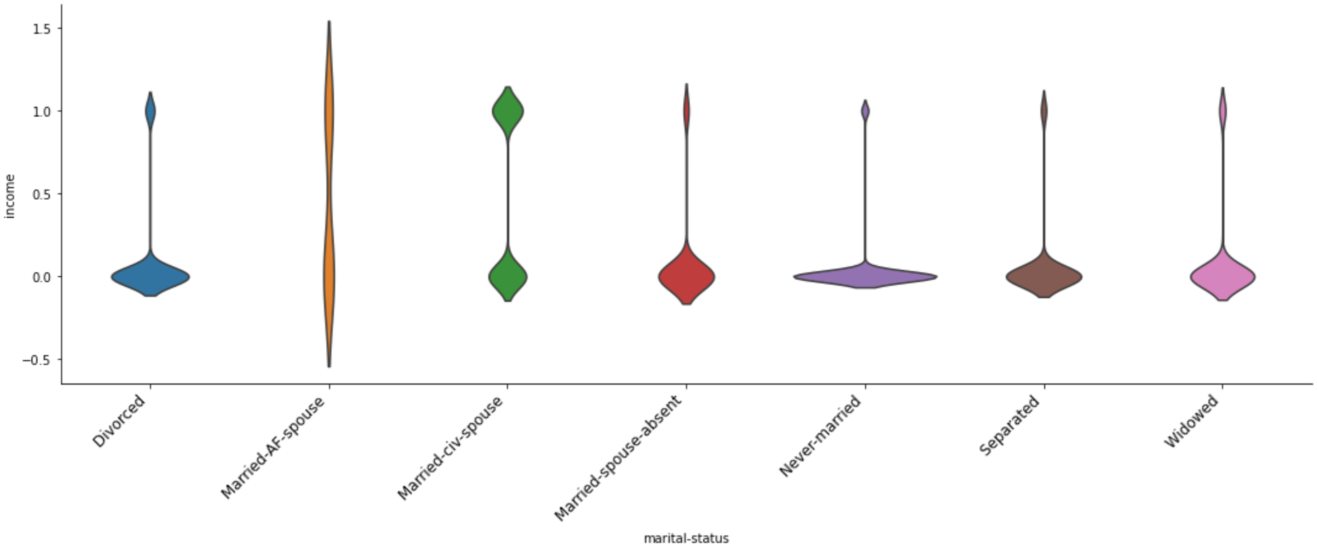
5. Education level clearly seems to influence income group as lower education categories have lower income and vice versa



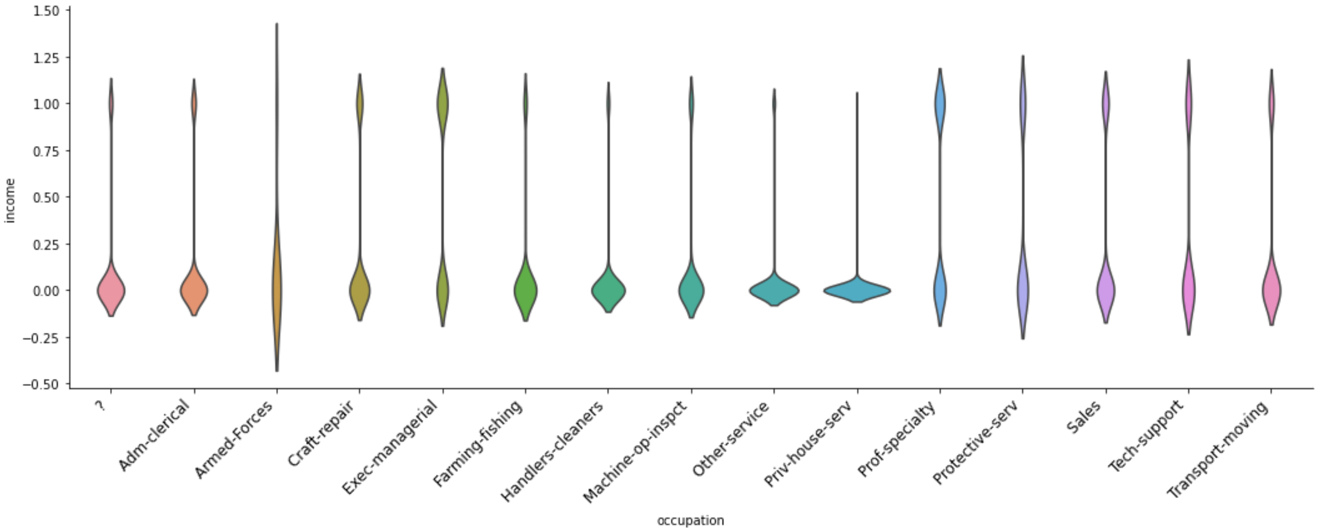
6. Workclass does not seem to impact income group to a great extent



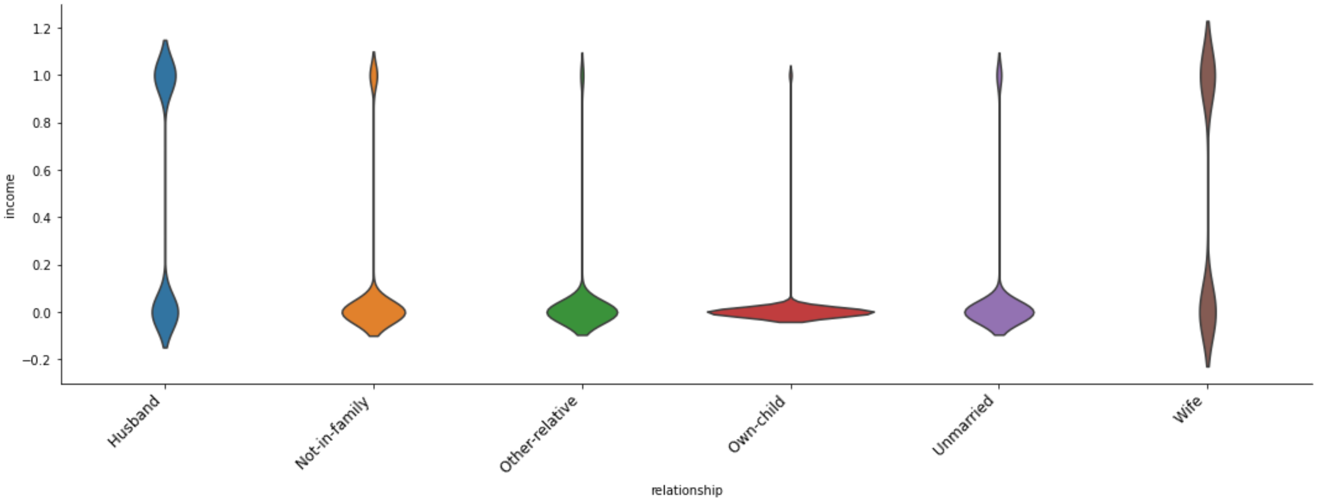
7. Marital-status does seem to have some impact on income groups. Specifically the 'Married with Civilian Spouse' group which clearly shows a higher proportion of distribution in the high income group. For the rest of the categories, the majority of the population lies in the low income group.



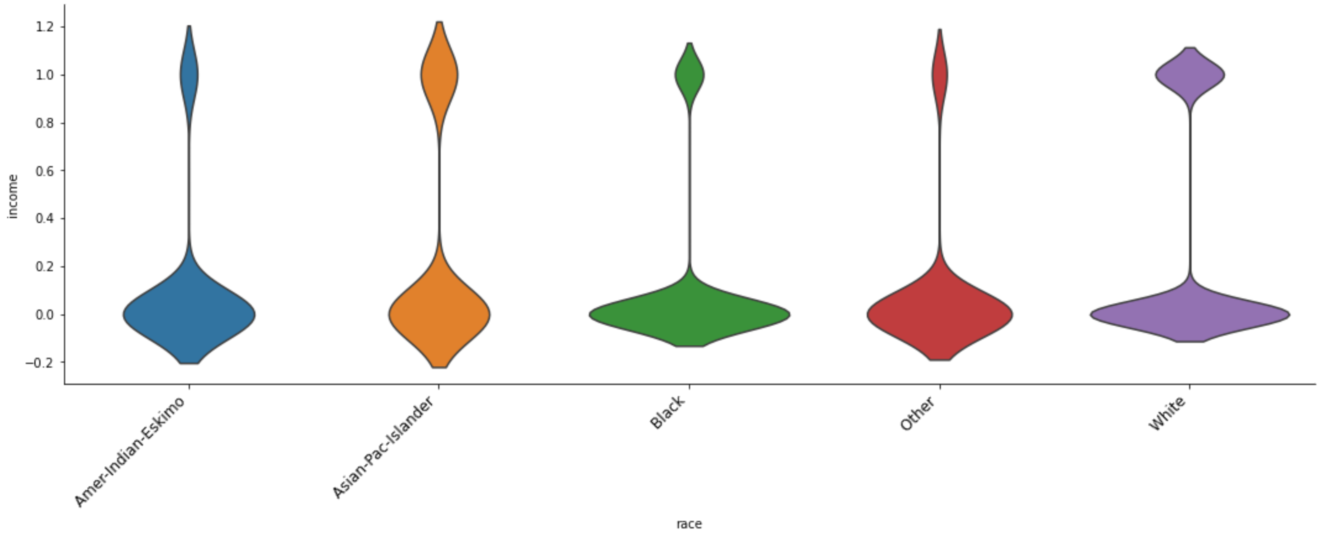
8. Following occupation show more proportion people falling in high income group - prof-speciality, protective services, Sales, Tech Support and Transport-moving



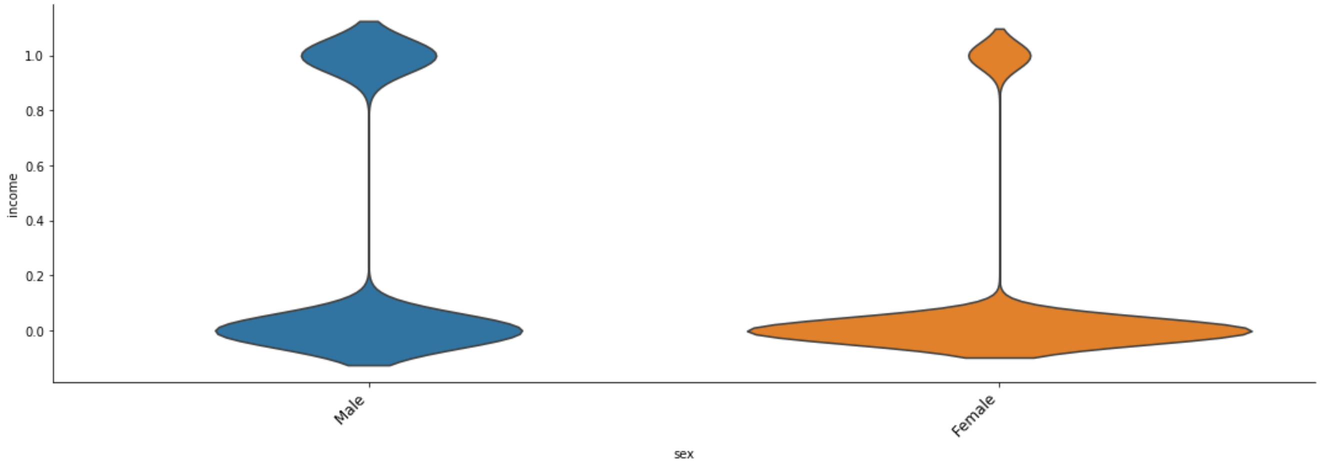
9. Husband-Wife category expectedly have higher income (possibly because they makeup the married with civilian wife group which has higher income as per the earlier chart)



10. There is more representation of white and Asian Pacific Islander among high income groups. Where all other races are more represented by lower income group



11. Females have a significantly lower representation in higher income group compared to Males



- **Closing Thoughts:**

As summarized above, many of the attributes have significant impact on affecting the income level of a person (e.g. age, hours-worked, profession, sex, education, etc.), while some of the attributes did not seem to have much impact on the income earned (e.g. workclass).

It remains to be seen however; whether this impact is statistically significant or not (i.e. the difference between two income groups is significantly substantial and is not occurring because of random sampling errors or internal variations of the data itself.

We will address this need in the next section of the project - Statistical Data Analysis using inferential statistical methods.

Further Analysis Using Statistical Analysis Techniques

We have already completed the Data Wrangling and Data Story part of the project. Now based on the exploratory data analysis done so far, we will first forms some hypothesis based on the results of EDA and then we will apply the inferential statistical principles and methods to test the validity of the hypothesis and finally forms some conclusions about the features of this dataset which should ultimately help us in building the requisite Machine Learning models to predict the prices effectively and accurately.

Additional Data Cleanup:

Based on the data analysis using value_counts method, I dropped following columns:

1. Fnlwgt: This is not a real attribute of the respondents and instead just stores unique IDs for each of the respondents
2. Capital-Gain and Capital-Loss: Majority of the respondents did not have capital-gains and capital loss reported in the survey (0 values). Since this would skew the sample, I dropped these 2 features from the dataset.

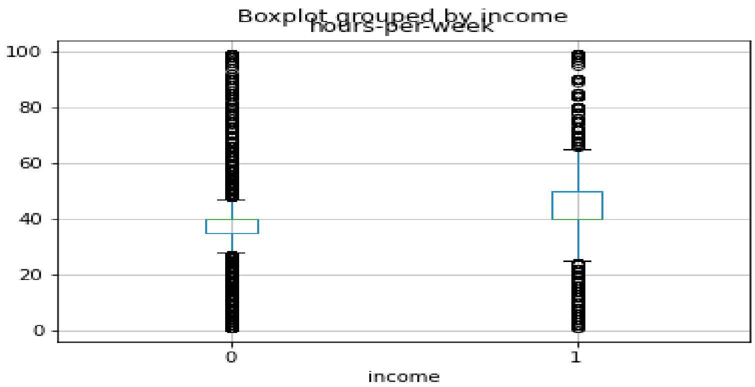
Data Transformations:

Following data transformations were used to facilitate ease of processing, apply some statistical and machine learning techniques using Pandas

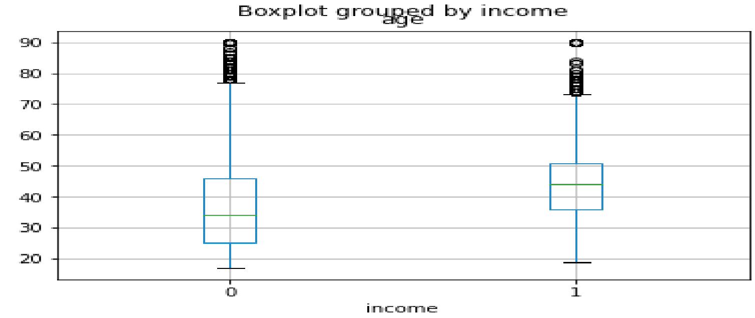
1. Change the data type of non-integer attributes from 'object' to 'category' for ease of processing. later. As a result the memory usage reduced from a previous value of 3.0+ MB to 1.2+ MB, a significant reduction indeed (60% reduction).
2. Change the values <=50K and >50K to 0 and 1 respectively, so that these values can then be utilized easily for EDA and ML modeling later.

Additional Observations based on further data analysis

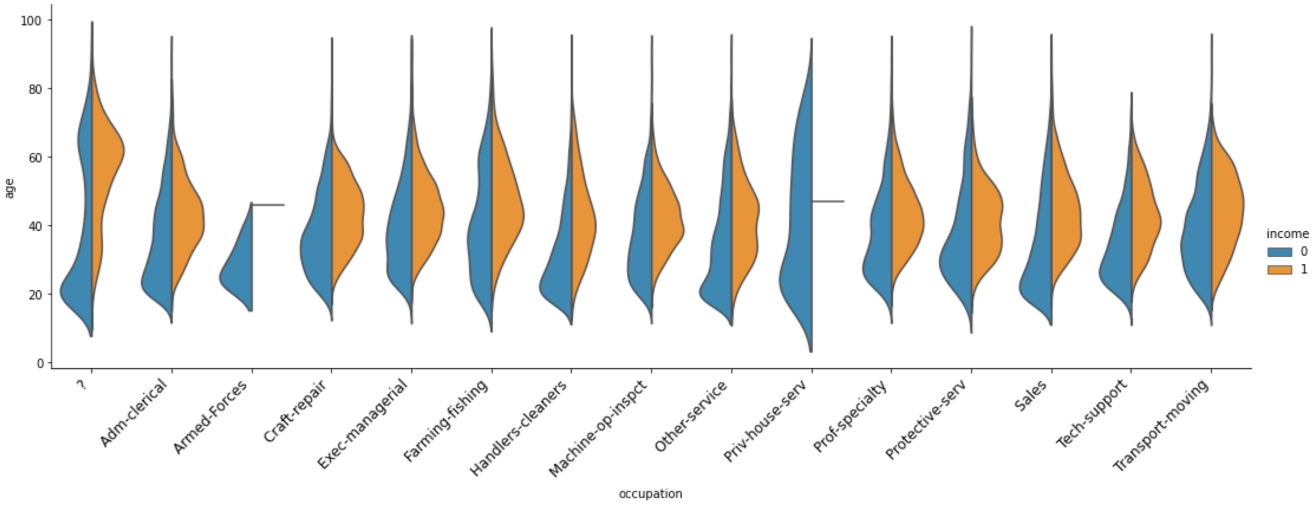
A. Out of a total sample of 32,561, majority of the population - 24,720 people (76%) falls in the low income, <=50K group, with rest of the population - 7841 people (24%) falling in high income, >50K group.
B. **Number of hours clocked by <= 50K and >50K group**: The >50K income group clearly has people who are clocking many more hours than people from <=50K income group. The average # of hours are more than 40 hours in high income group vs. an average of less than 40 hours in the low income group.



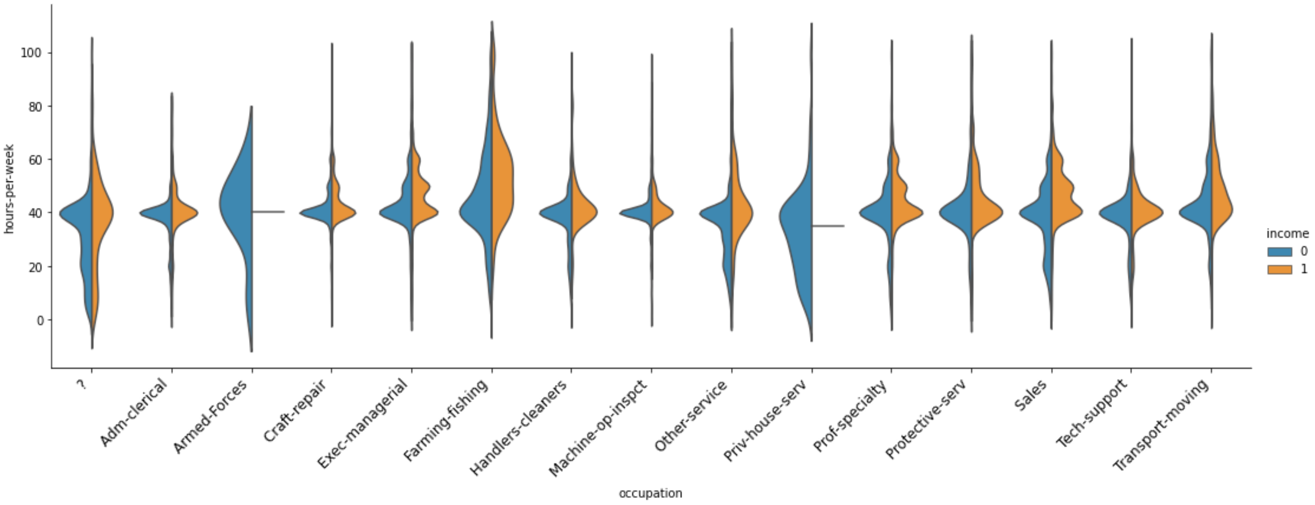
C. **Age distribution among <= 50K and >50K segments**: The >50K income group clearly has people with higher age than people from <=50K income group. The average age is 44.25 years in high income group vis-a-vis average age of 36.87 years in low income group



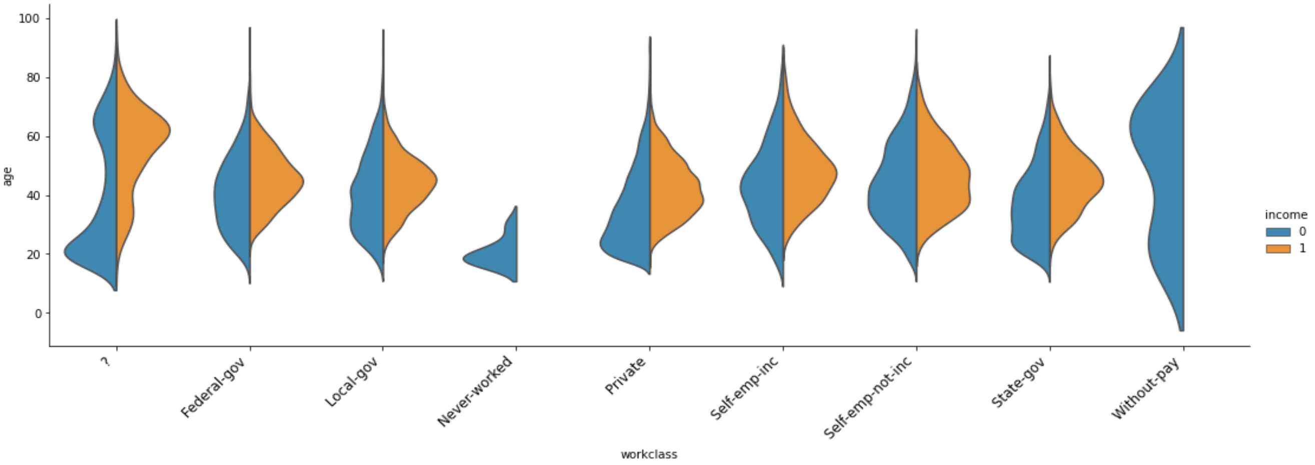
D. The overall trend of age difference between higher income group and lower income group continues across various occupation categories as well.



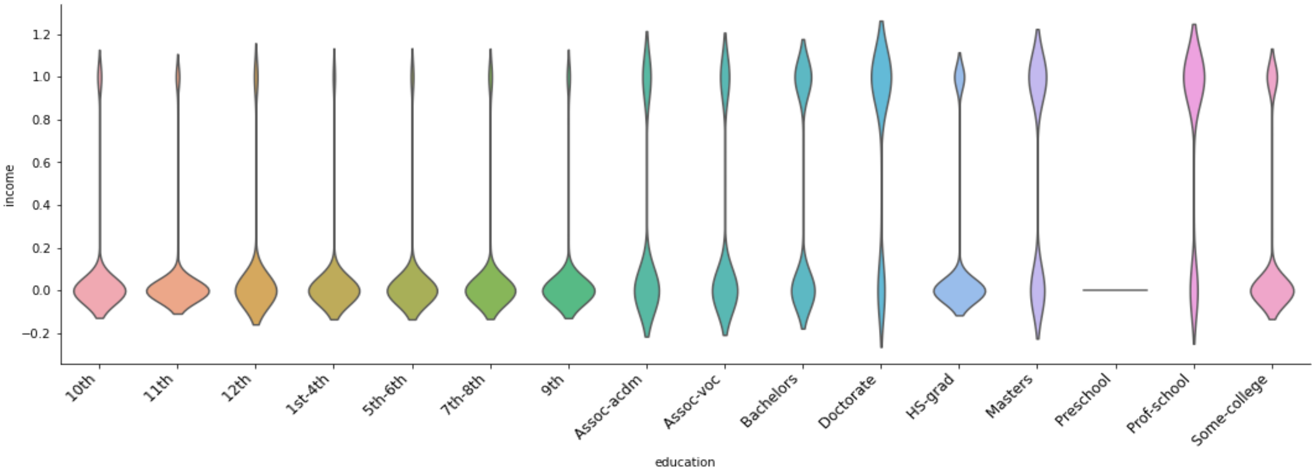
E. The overall trend of 'hours worked per week' difference between higher income group and lower income group continues across various occupation categories as well.



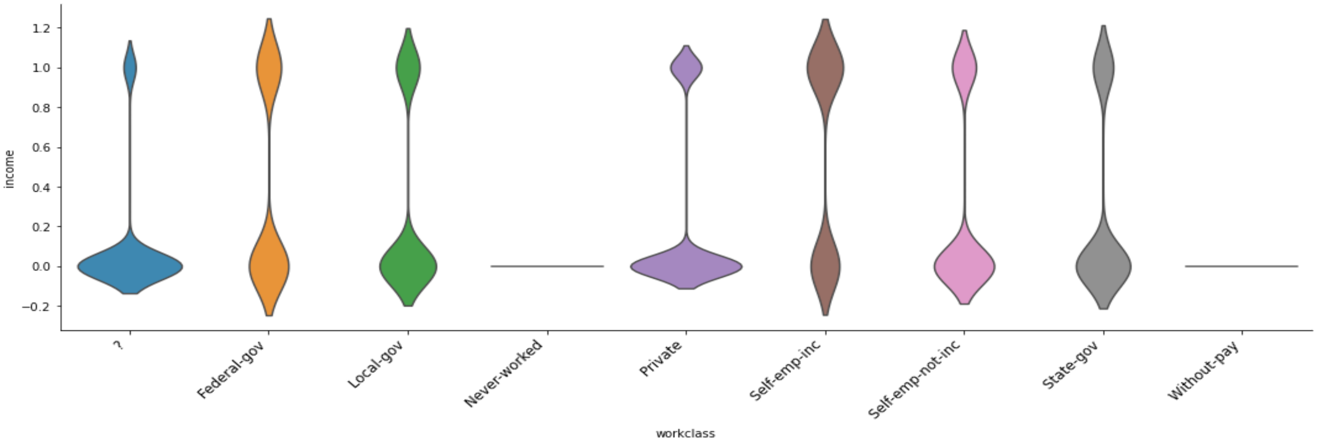
F. Across the work classes, for the higher income group, the age is more skewed towards higher values. The overall age difference between the low income group and high income group seems to be substantial. Lower income group is more concentrated between the 20 to 40 years age block vs. the higher income group which is more concentrated between the 40 to 60 age block.



G. Education level clearly seems to influence income groups as lower education categories have lower income and vice versa.



H. Workclass does not seem to impact income groups to a great extent.



I. Findings from further EDW
   a. For attribute 'native-country' the dominating value is USA
   b. For attribute 'hours-per-week' the dominating value is @40 hours
   c. For the attribute 'race' the dominating value is for White race. For these attributes, later on we will see if we can aggregate the 'other' values in one attribute (e.g. for native-country we could divide the dataset between 'USA' and 'NonUSA')
   d. For all other attributes the data is spread across many categories. Thus we will have to do further analysis for all these categorical attributes to find out their influence on income.

Statistical Analysis

Let's do the analysis of how attributes age, education-num, hours-per-week differ between people with <=50K and >50K income.

Age:

- Mean of age of people with low income is 36.78373786407767
- Standard Deviation of age of people with low income is 14.019804910115214

Let's state the null and alternative hypothesis here. Use the t-test for the difference between means where the pooled standard deviation of the two groups is given by

Null Hypothesis : There is no significant difference between the age distribution between people with high income vs. people with low income.

Alternate Hypothesis: There is a significant difference between the age distribution between people with high income vs. people with low income.

The 't-value' as calculated by the 3 methods above (t1 = 43.44, t2 = 43.44, t3= 50.27) is significantly higher compared to the mean value of the two groups (mean age value of high income group = 44.25 and mean age value of low income group = 36.78). Also the p-value (the probability of t value being so high due to the random sampling error is given as 0.00), which establishes the fact that the difference in the means is significant.

Considering the results above, we reject the Null Hypothesis that there is no significant difference between the age distribution of 2 groups and accept the Alternate Hypothesis that there is indeed a significant difference between the age distribution of High Income and Low Income groups.

It is clear that there is a significant difference in the age distribution between the high income group and low income group. It is unlikely that this difference has occured because of sampling or experimental error.

Education:

The 't-value' as calculated by the 3 methods above (t1 = 64.19, t2 = 64.19, t3= 64.90) is significantly higher compared to the mean value of the two groups (mean education-num value of high income group = 11.61 and mean education-num value of low income group = 9.59). Also the p-value (the probability of t value being so high due to the random sampling error is given as 0.00), which establishes the fact that the difference in the means is significant.

Considering the results above, we reject the Null Hypothesis that there is no significant difference between the education distribution of 2 groups and accept the Alternate Hypothesis that there is indeed a significant difference between the education distribution of High Income and Low Income groups.

It is clear that there is a significant difference in the education distribution between the high income group and low income group. It is unlikely that this difference has occured because of sampling or experimental error.

Hours-per-Week:

The 't-value' as calculated by the 3 methods above (t1 = 42.59, t2 = 42.59, t3= 45.13) is significantly higher compared to the mean value of the two groups (mean 'hours-per-week' value of high income group = 45.47 and mean 'hours-per-week' value of low income group = 38.84). Also the p-value (the probability of t value being so high due to the random sampling error is given as 0.00), which establishes the fact that the difference in the means is significant.

Considering the results above, we reject the Null Hypothesis that there is no significant difference between the 'hours-per-week' distribution of 2 groups and accept the Alternate Hypothesis that there is indeed a significant difference between the 'hours-per-week' distribution of High Income and Low Income groups.

It is clear that there is a significant difference in the 'hours-per-week' distribution between the high income group and low income group. It is unlikely that this difference has occured because of sampling or experimental error.

With details on statistical analysis of the 3 attributes (age, education, hours-per-week) as given above, we would like to close this part of capstone project 1.

In the next section we will work on applying the machine learning principals, concepts and models to arrive at a model which can be used to predict the income of the sample based on the demographic data as collected in the sensus.

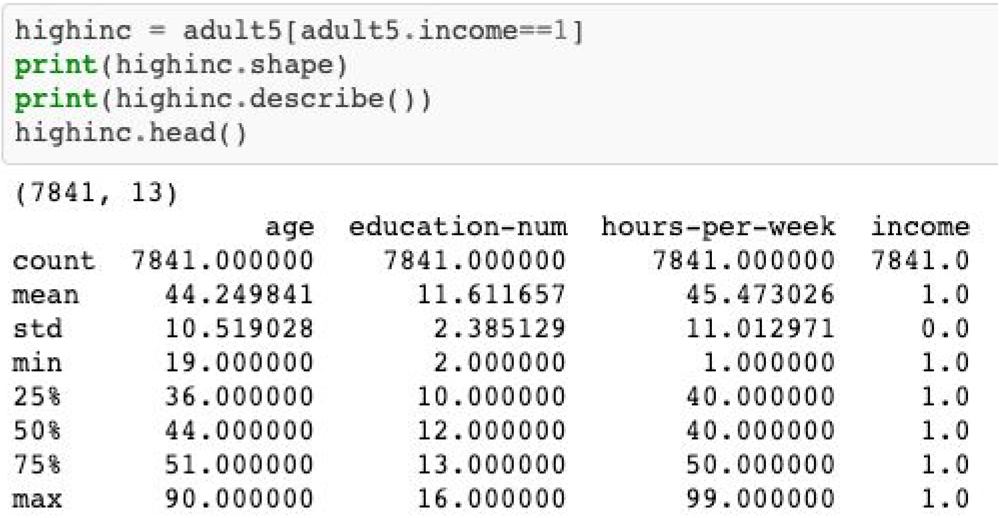
Further Analysis Using Machine Learning Techniques

We have already completed the Data Wrangling and Data Story part of the project. In the statistical analysis section, we did form some hypotheses based on the results of EDA and tested the validity of those hypotheses by applying the inferential statistical principles and methods. Based on the conclusions about the features of this dataset, now let’s build the requisite Machine Learning models to predict the prices effectively and accurately.

Segmenting (Splitting) Sample in 2 Segments <=50K and >50K

Let's do the analysis of how attributes age, education-num, hours-per-week differ between people with <=50K and >50K income.

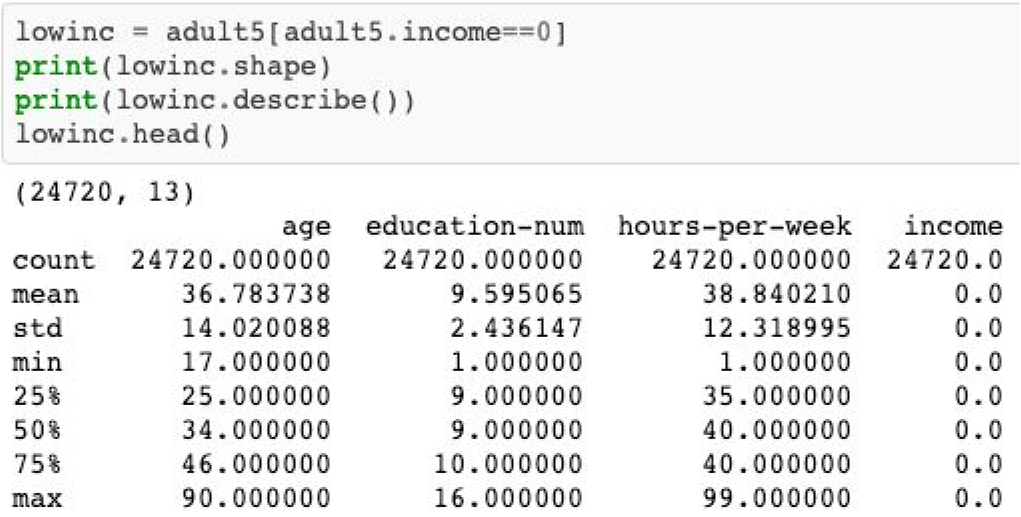
Summary description of High Income (>50K income) Group:



```
highinc = adult5[adult5.income==1]
print(highinc.shape)
print(highinc.describe())
highinc.head()
```

```
(7841, 13)
               age  education-num  hours-per-week  income
count  7841.000000    7841.000000     7841.000000  7841.0
mean     44.249841      11.611657       45.473026     1.0
std      10.519028       2.385129       11.012971     0.0
min      19.000000       2.000000        1.000000     1.0
25%      36.000000      10.000000       40.000000     1.0
50%      44.000000      12.000000       40.000000     1.0
75%      51.000000      13.000000       50.000000     1.0
max      90.000000      16.000000       99.000000     1.0
```

Summary Description of Low Income (<=50K income) Group:

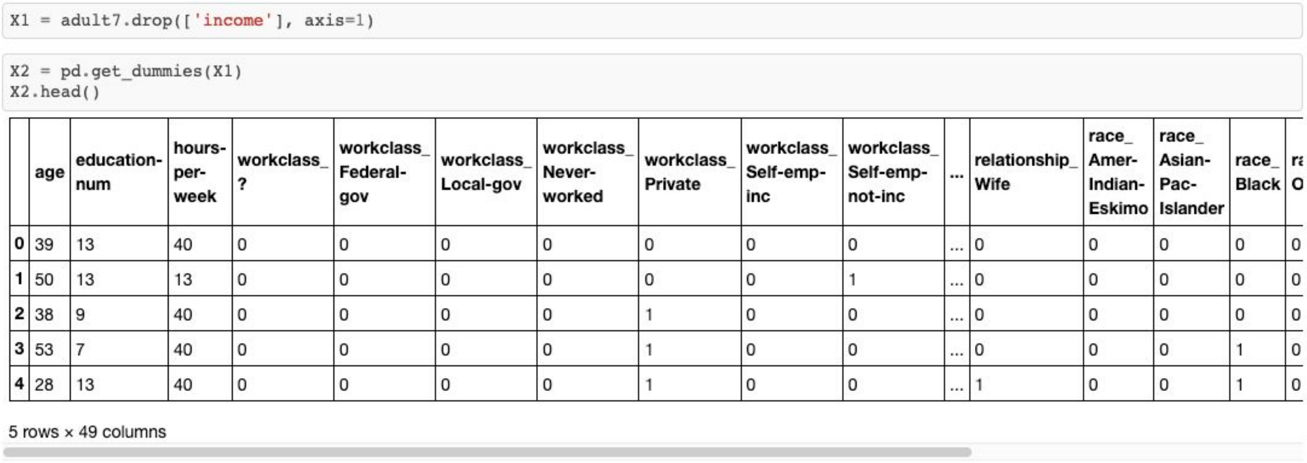


```
lowinc = adult5[adult5.income==0]
print(lowinc.shape)
print(lowinc.describe())
lowinc.head()
```

```
(24720, 13)
                age  education-num  hours-per-week   income
count  24720.000000   24720.000000    24720.000000  24720.0
mean      36.783738       9.595065       38.840210      0.0
std       14.020088       2.436147       12.318995      0.0
min       17.000000       1.000000        1.000000      0.0
25%       25.000000       9.000000       35.000000      0.0
50%       34.000000       9.000000       40.000000      0.0
75%       46.000000      10.000000       40.000000      0.0
max       90.000000      16.000000       99.000000      0.0
```

You can visually inspect the difference between the 2 income groups based on the statistical descriptions of the two groups as shown above.

Changing Categorical Features in to Continuous Integers type by using Pandas get_dummies function



```
X1 = adult7.drop(['income'], axis=1)
```

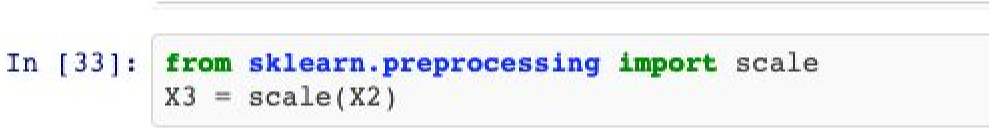
```
X2 = pd.get_dummies(X1)
X2.head()
```

	age	education-num	hours-per-week	workclass_?	workclass_Federal-gov	workclass_Local-gov	workclass_Never-worked	workclass_Private	workclass_Self-emp-inc	workclass_Self-emp-not-inc	...	relationship_Wife	race_Amer-Indian-Eskimo	race_Asian-Pac-Islander	race_Black	ra O
0	39	13	40	0	0	0	0	0	0	0	...	0	0	0	0	0
1	50	13	13	0	0	0	0	0	0	1	...	0	0	0	0	0
2	38	9	40	0	0	0	0	1	0	0	...	0	0	0	0	0
3	53	7	40	0	0	0	0	1	0	0	...	0	0	0	1	0
4	28	13	40	0	0	0	0	1	0	0	...	1	0	0	1	0

5 rows × 49 columns

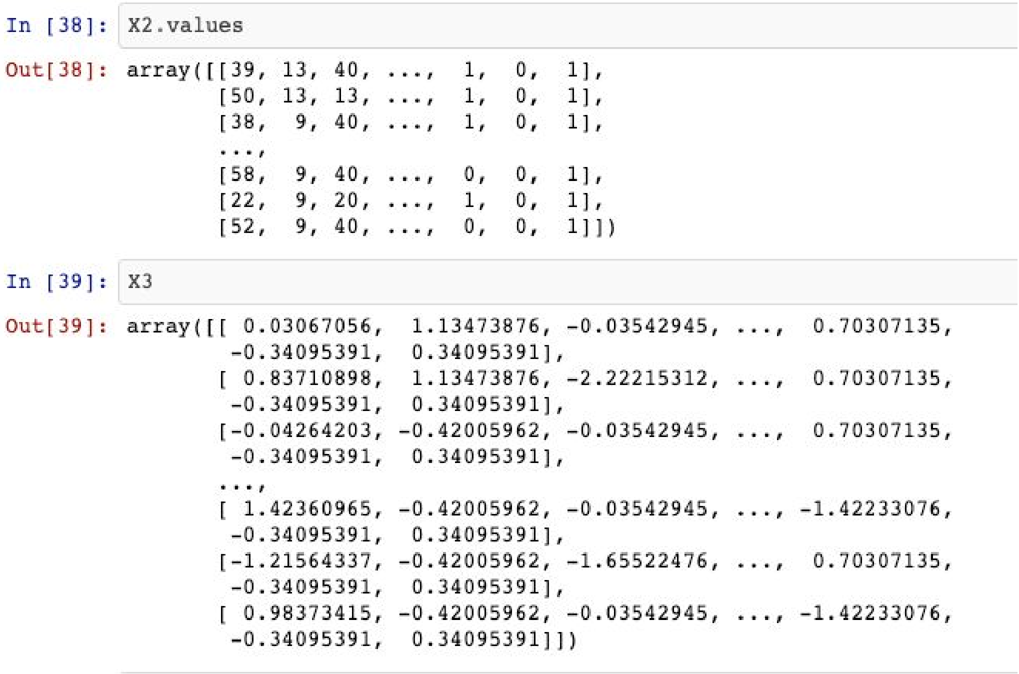
As you can see above, the original data frame with 12 features was transformed into a data frame with 49 features.

Now let's apply some scaling in order to ensure the various attributes are scaled adequately in proportion to each other. We will use the 'scale' class from sklearn.preprocessing.



```
In [33]: from sklearn.preprocessing import scale
         X3 = scale(X2)
```

Below you can see how the composition of arrays change after applying scaling:

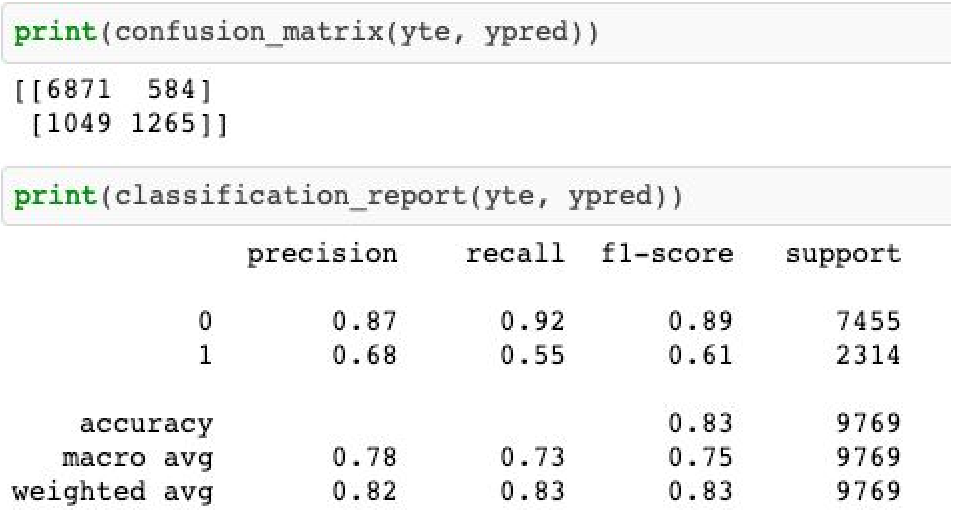


```
In [38]: X2.values

Out[38]: array([[39, 13, 40, ...,  1,  0,  1],
                [50, 13, 13, ...,  1,  0,  1],
                [38,  9, 40, ...,  1,  0,  1],
                ...,
                [58,  9, 40, ...,  0,  0,  1],
                [22,  9, 20, ...,  1,  0,  1],
                [52,  9, 40, ...,  0,  0,  1]])

In [39]: X3

Out[39]: array([[ 0.03067056,  1.13473876, -0.03542945, ...,  0.70307135,
                 -0.34095391,  0.34095391],
                [ 0.83710898,  1.13473876, -2.22215312, ...,  0.70307135,
                 -0.34095391,  0.34095391],
                [-0.04264203, -0.42005962, -0.03542945, ...,  0.70307135,
                 -0.34095391,  0.34095391],
                ...,
                [ 1.42360965, -0.42005962, -0.03542945, ..., -1.42233076,
                 -0.34095391,  0.34095391],
                [-1.21564337, -0.42005962, -1.65522476, ...,  0.70307135,
                 -0.34095391,  0.34095391],
                [ 0.98373415, -0.42005962, -0.03542945, ..., -1.42233076,
                 -0.34095391,  0.34095391]])
```

Now Let’s Apply Logistics Regression Model

Since this is a Classification problem, we will use the Logistic Regression model from SKLearn

Following are the results.

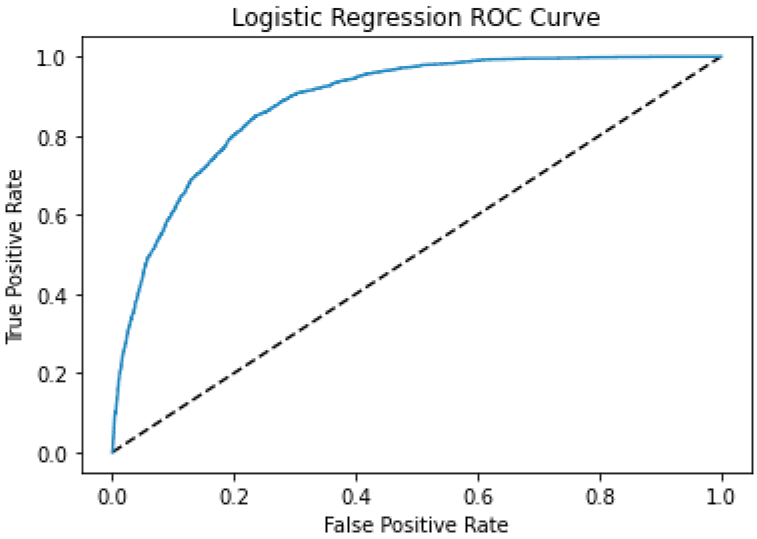


```
print(confusion_matrix(yte, ypred))
```

```
[[6871  584]
 [1049 1265]]
```

```
print(classification_report(yte, ypred))
```

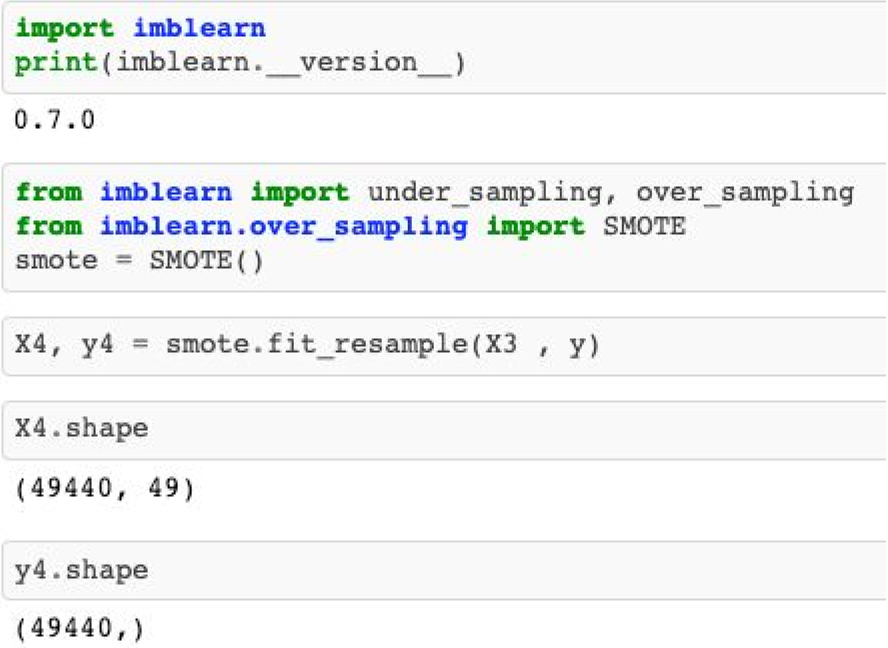
```
              precision    recall  f1-score   support

           0       0.87      0.92      0.89      7455
           1       0.68      0.55      0.61      2314

    accuracy                           0.83      9769
   macro avg       0.78      0.73      0.75      9769
weighted avg       0.82      0.83      0.83      9769
```



As you can see above, since the area under the ROC curve is large, we can safely assume that the model used is a good model to use for predicting the income class.

Oversampling using SMOT (Synthetic Minority Oversampling Technique):

Since the >50K class is under-sampled the precision, recall and f1-score of this class is not great. This has also impacted the overall scores of the results. This can be corrected by oversampling the under-represented/minority class - >50K income class in this case. Let's use the SMOT class to oversample the minority class and rerun the Logistic Regression model on the new sample.



```
import imblearn
print(imblearn.__version__)
```

```
0.7.0
```

```
from imblearn import under_sampling, over_sampling
from imblearn.over_sampling import SMOTE
smote = SMOTE()
```

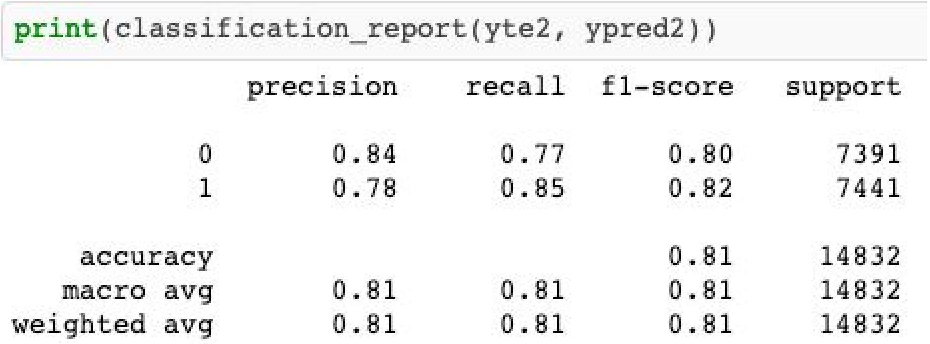
```
X4, y4 = smote.fit_resample(X3 , y)
```

```
X4.shape
```

```
(49440, 49)
```

```
y4.shape
```

```
(49440,)
```



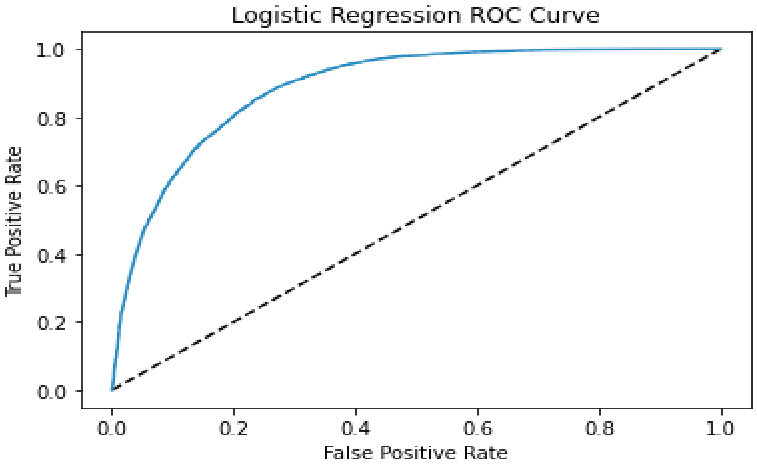
```
print(classification_report(yte2, ypred2))
```

```
              precision    recall  f1-score   support

           0       0.84      0.77      0.80      7391
           1       0.78      0.85      0.82      7441

    accuracy                           0.81     14832
   macro avg       0.81      0.81      0.81     14832
weighted avg       0.81      0.81      0.81     14832
```

Key findings based on Classification Report after applying SMOT Technique

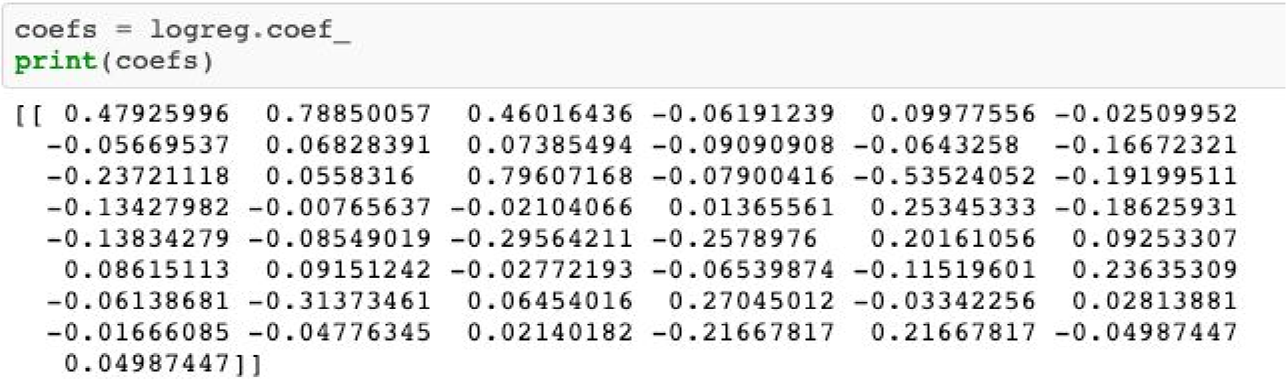
As you can see above the precision, recall and f1 score for class 1 (>50K income class) have gone up significantly with the help of oversampling using SMOT technique. However also note that the overall score (weighted average) did not improve much, and in fact reduced by 1 percentage point compared to the results without oversampling.

However we would still prefer these scores over the one without oversampling because the second sample has better and balanced representation of both the <=50K and >50K income groups



As you can see above, since the area under the ROC curve is large, we can safely assume that the model used is a good model to use for predicting the income class.

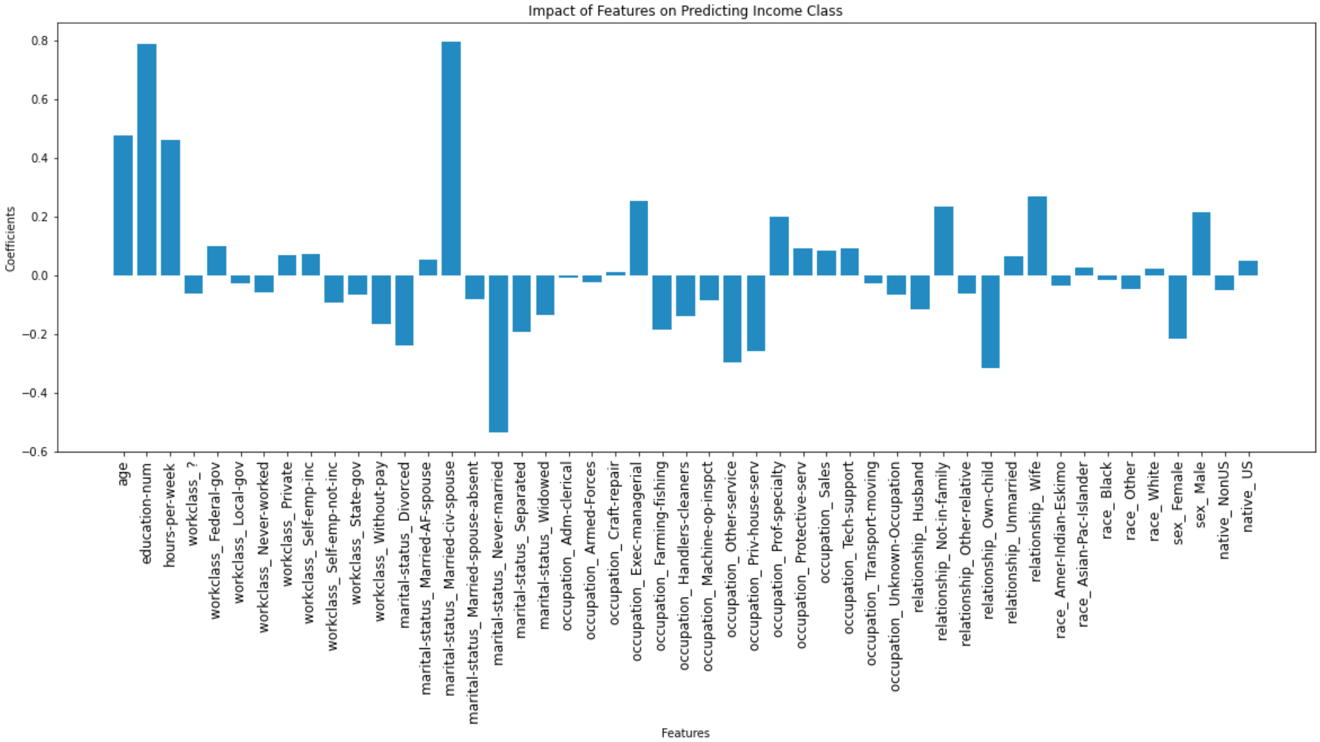
Which Features Have High/Significant Impact on Determining/Predicting Income?
We can identify which features have high/significant impact by analyzing the coefficient values.



```
coefs = logreg.coef_
print(coefs)
```

```
[[ 0.47925996  0.78850057  0.46016436 -0.06191239  0.09977556 -0.02509952
  -0.05669537  0.06828391  0.07385494 -0.09090908 -0.0643258  -0.16672321
  -0.23721118  0.0558316   0.79607168 -0.07900416 -0.53524052 -0.19199511
  -0.13427982 -0.00765637 -0.02104066  0.01365561  0.25345333 -0.18625931
  -0.13834279 -0.08549019 -0.29564211 -0.2578976   0.20161056  0.09253307
   0.08615113  0.09151242 -0.02772193 -0.06539874 -0.11519601  0.23635309
  -0.06138681 -0.31373461  0.06454016  0.27045012 -0.03342256  0.02813881
  -0.01666085 -0.04776345  0.02140182 -0.21667817  0.21667817 -0.04987447
   0.04987447]]
```

Now let's plot these coefficient values for each of the attributes in a graph



Key findings based on analysis of coefficient values for features

1. Age, Education, Hours-per-Week, Married W Civil Souse, Executive-Managerial Occupation, Male categories have high and positive impact on determining/predicting income of individuals

2. Marital Status of Never Married, Relationship of Own Child have significant but negative correlation with income

3. Male class is positively correlated with income while Female class is negatively correlated with income (this could be because of lower employment status among female class

4. Race does not seem to have any impact on predicting the income class

5. Many low skilled income categories (e.g. Farming, Fishing, Handlers, Cleaners) have a negative correlation with income.

Closing Thoughts, Conclusions

- Two income groups, High Income (>50K) and Low Income (<=50K) have distinct characteristics as defined by the features of the two groups
- In order to use Statistical and Machine Learning techniques, one would need to do quite a bit of data cleanup (changing data types from Objects, Categories, Integers, etc,), EDA, data transformations (e.g. scaling the data, dropping columns, splitting samples, etc.)
- Machine Learning models can be applied more effectively by first transforming the data to reduce impact of inherent sample structure imbalances (e.g. small number of samples for high income group). We used data scaling, converting categorical features into integer types via using get_dummies method and oversampling using SMOT.
- **Age, Education, Hours-per-Week, Married W Civil Souse, Executive-Managerial Occupation, Male categories have high and positive impact on determining/predicting income of individuals.**