Springboard Data Science - Capstone Project 1 - Milestone Report

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:

Listing of attributes

- 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

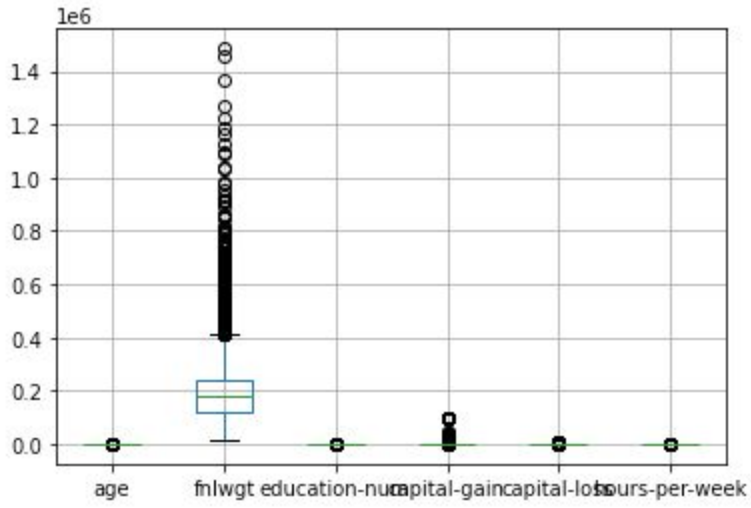
the later statistical analysis. If required, I can drop these rows later in the EDA and Data Story part of the project.

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>
```



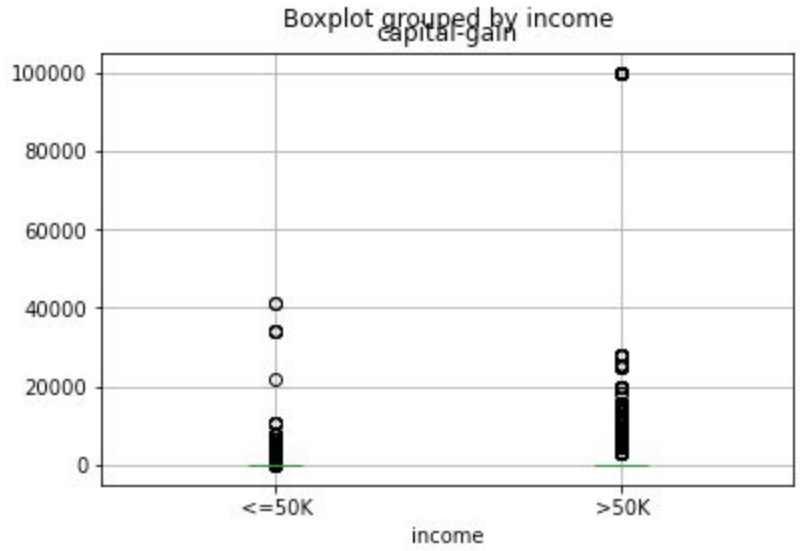
The above box-plot cannot be used for EDA since the data is not normalized and as a result Box-plot cannot be displayed in a useful way. I will plot individual columns against each other to get more interpretable results of box-plots.

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.

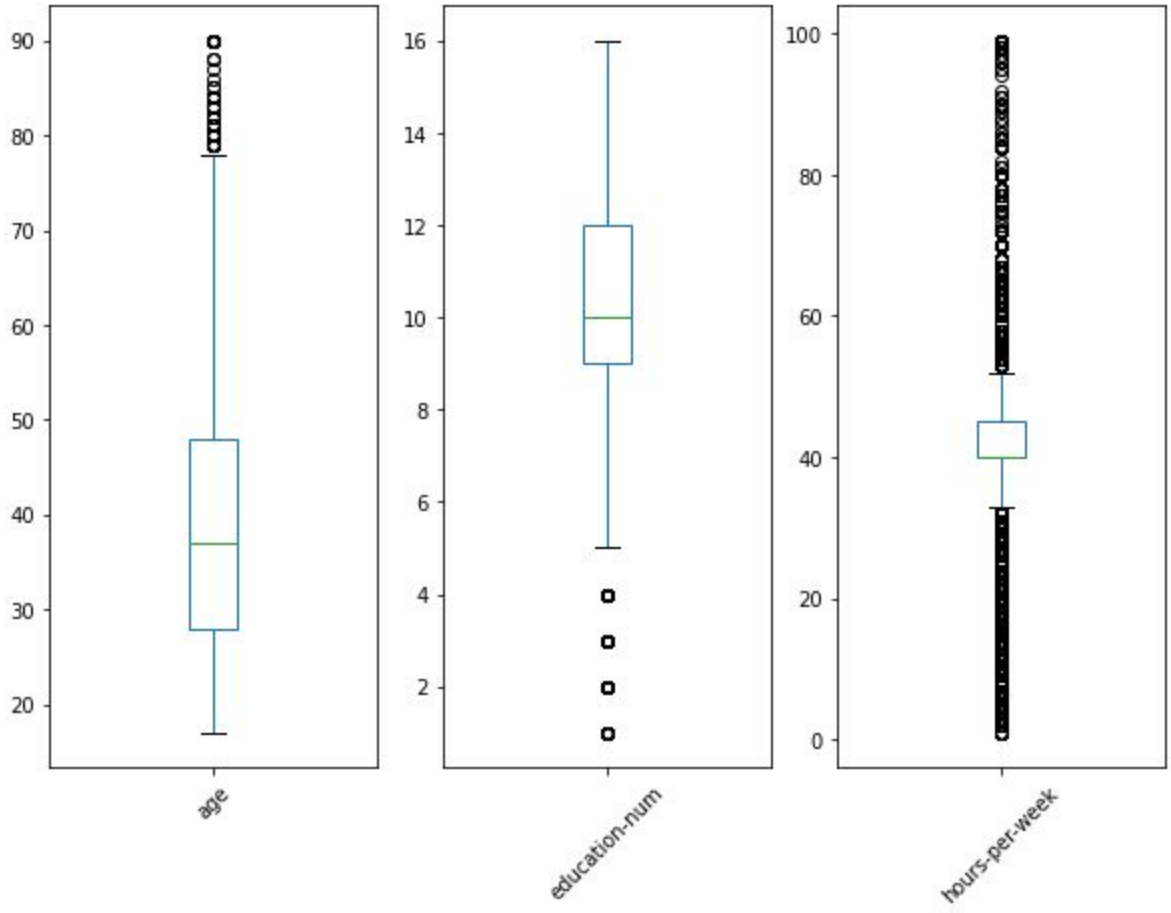
```
adult.describe()
```

	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.

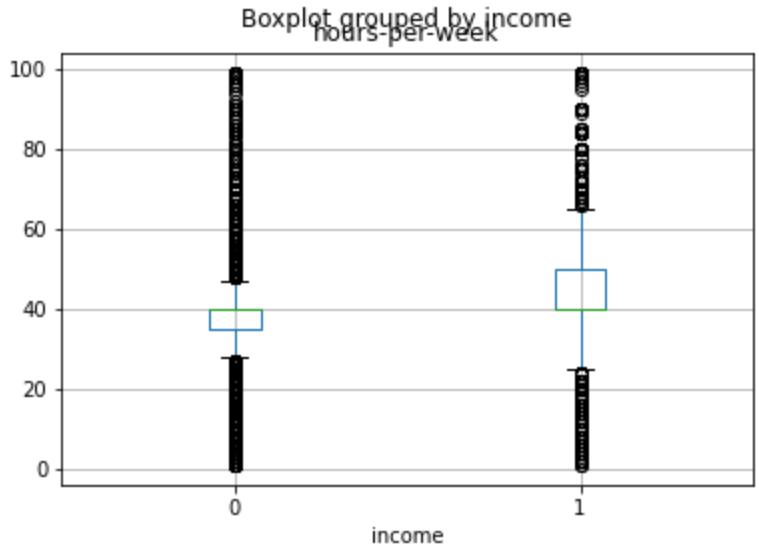
The Income attribute values which are either <=50K and >50K, were converted (transformed) to 0 and 1 (and integer type), so that this attribute can be used for further statistical analysis.

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

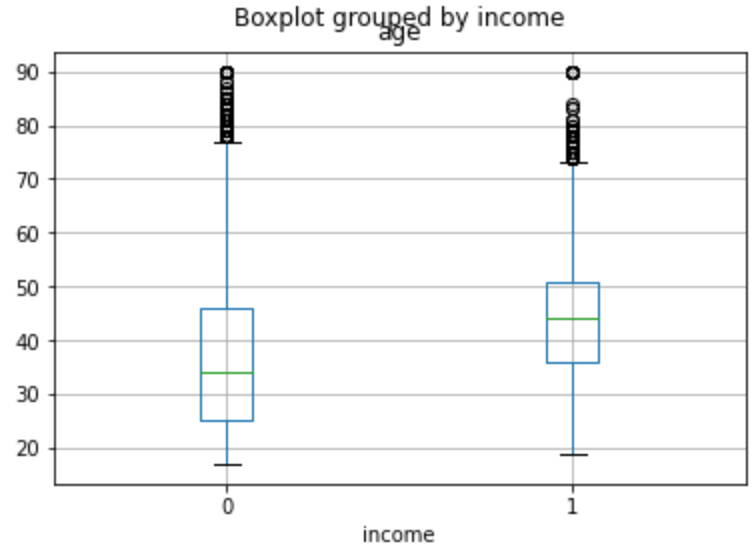
	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	hours-per-week	native-country	income
0	39	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	40	United-States	0
1	50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	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:

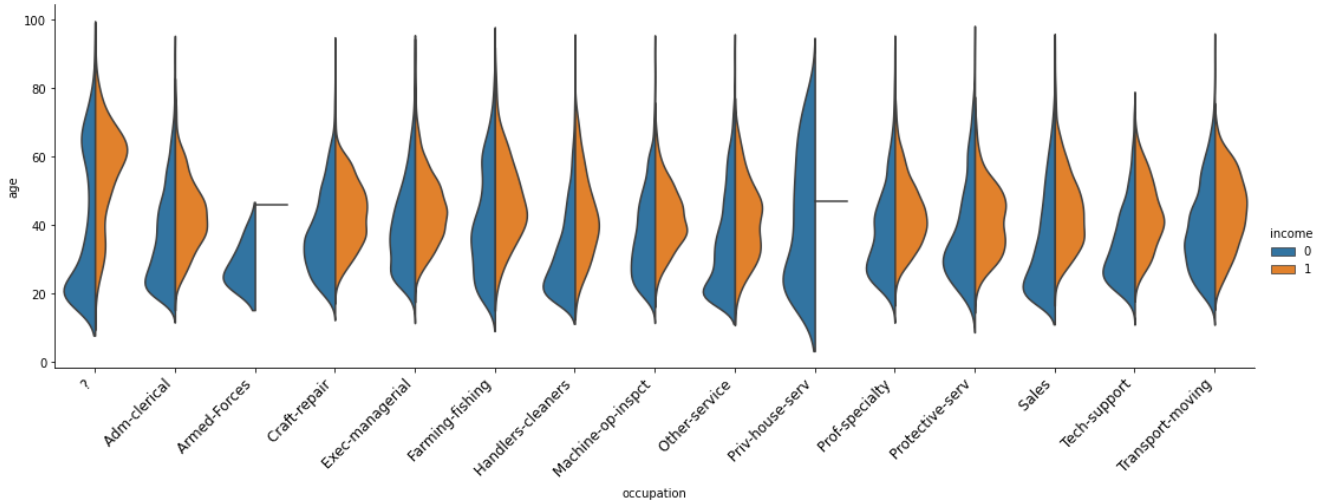
1. The low income group and high income group clearly differ from each other based on the data for attributes age and hours-per-week.
    a. The >50K income group clearly has people who are clocking many more hours than people from <=50K income group. The average # of hours are more than 40 hours in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



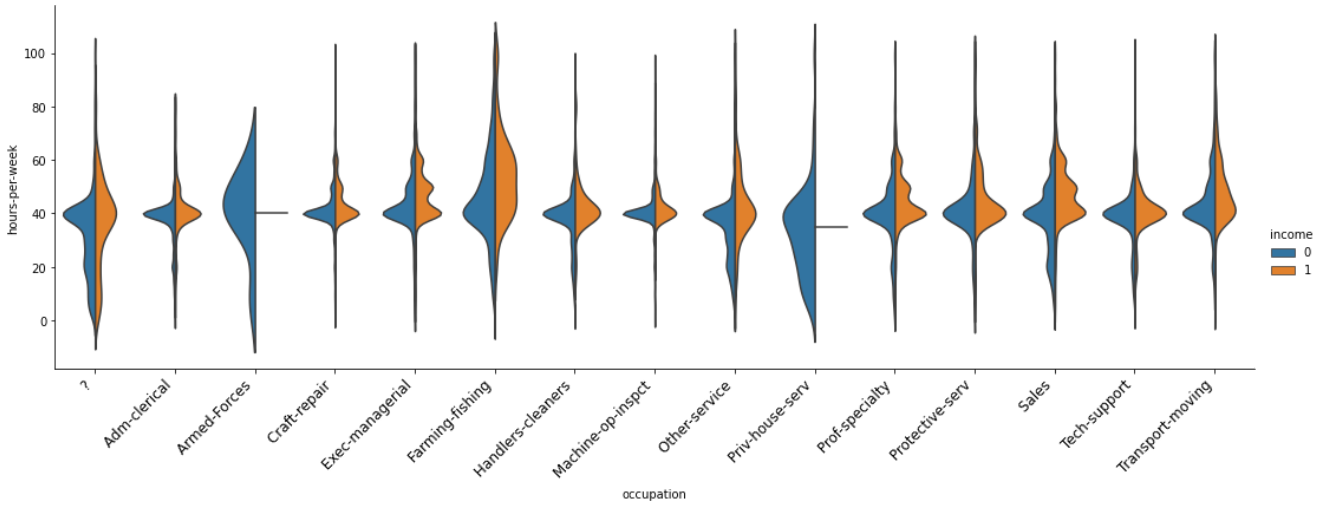
    b. The >50K income group clearly has people with higher age than people from <=50K income group. The average age is 44.25 years in high income group vis-a-vis average age of 36.87 years in low income group



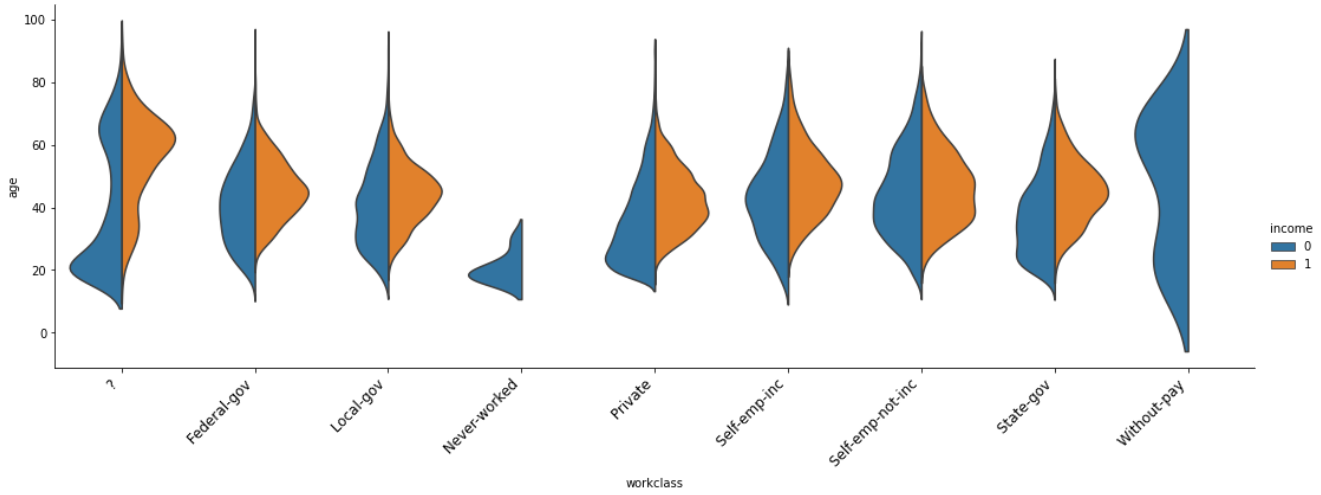
2. The overall trend of age difference between higher income group and lower income group continues across various occupation categories as well



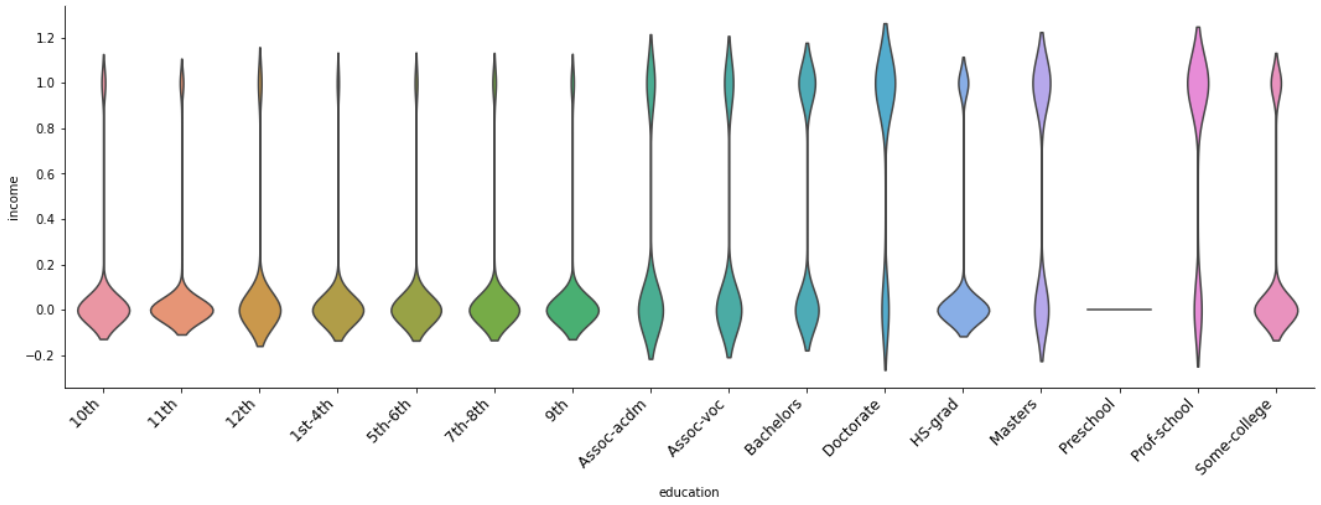
3. The overall trend of 'hours worked per week' difference between higher income group and lower income group continues across various occupation categories as well



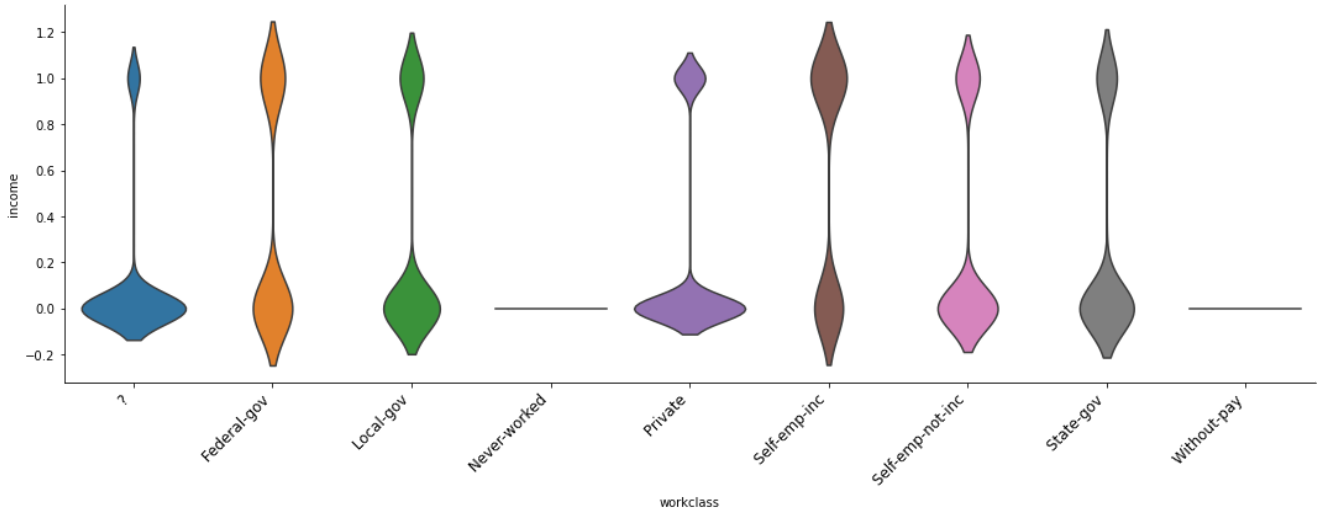
4. Across the work classes, for the higher income group, the age is more skewed towards higher values. The overall age difference between the low income group and high income group seems to be substantial. Lower income group is more concentrated between 20 to 40 years age block vs. higher income group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



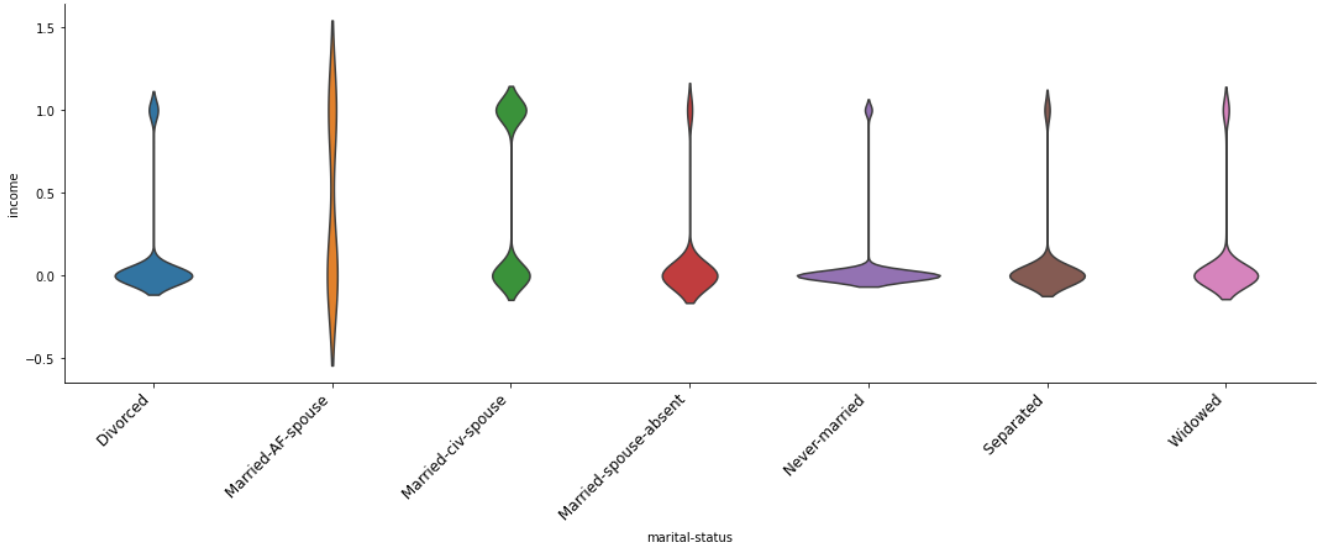
5. Education level clearly seems to influence income group as lower education categories have lower income and vice versa



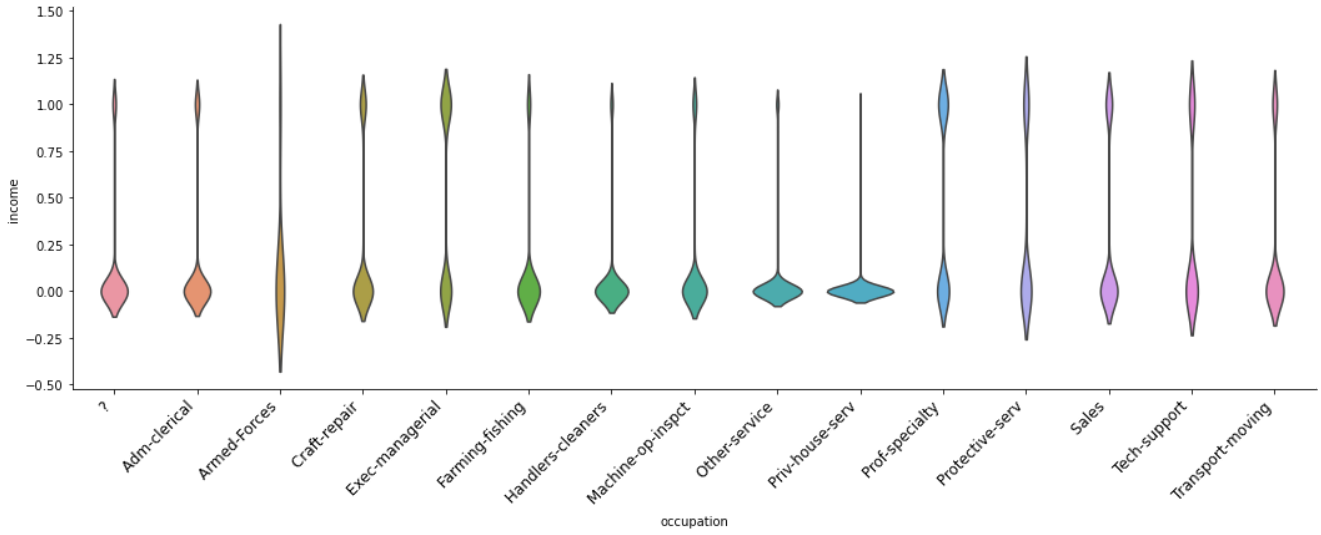
6. Workclass does not seem to impact income group to a great extent



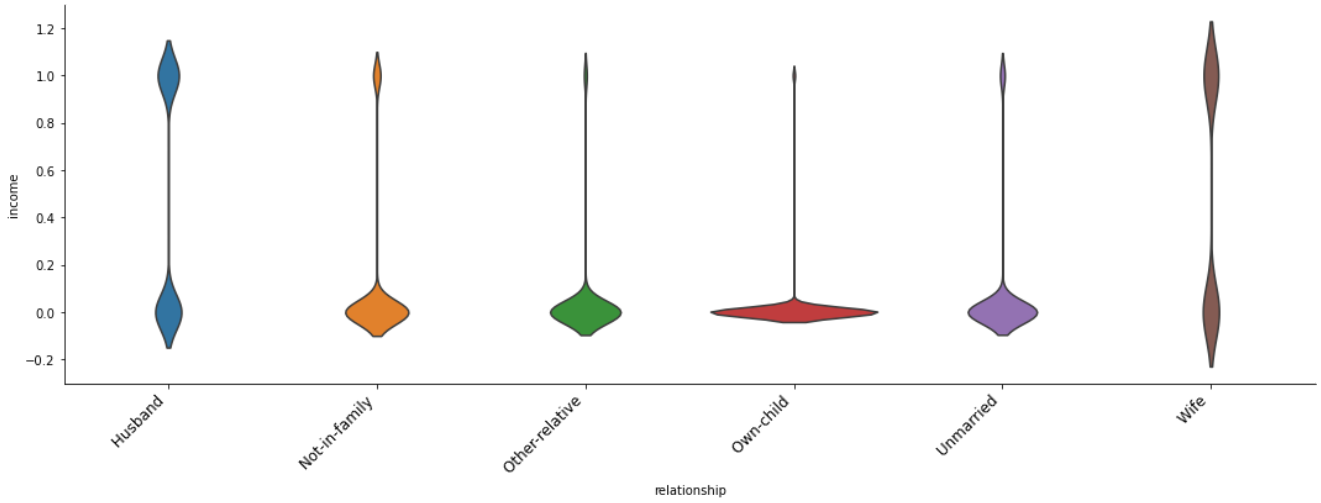
7. Marital-status does seem to have some impact on income groups. Specifically the 'Married with Civilian Spouse' group which clearly shows a higher proportion of distribution in the high income group. For the rest of the categories, the majority of the population lies in the low income group.



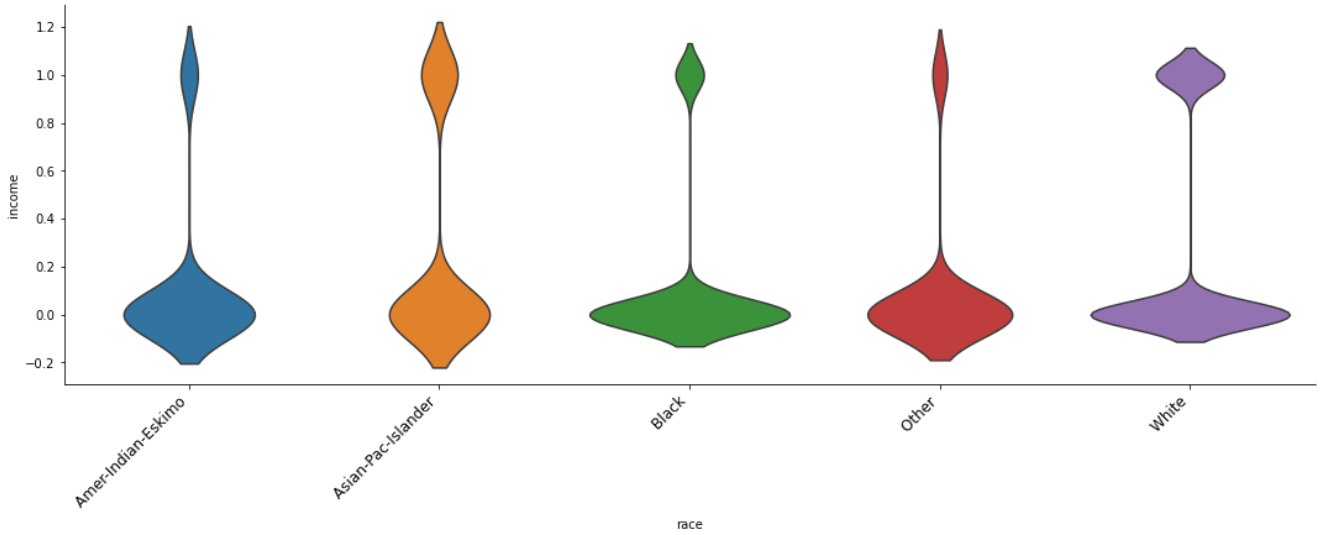
8. Following occupation show more proportion people falling in high income group - prof-speciality, protective services, Sales, Tech Support and Transport-moving



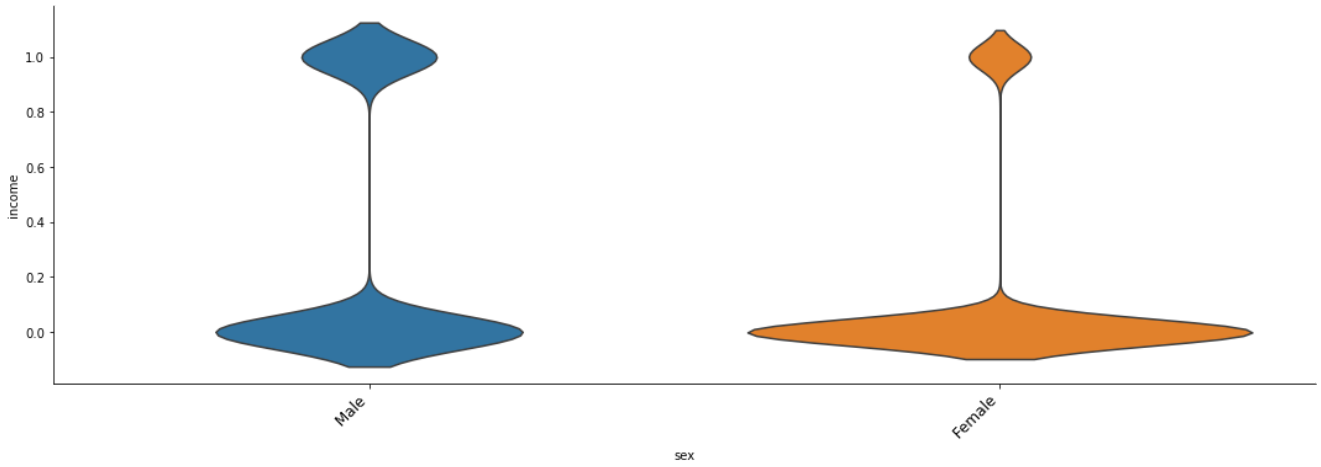
9. Husband-Wife category expectedly have higher income (possibly because they makeup the married with civilian wife group which has higher income as per the earlier chart)



10. There is more representation of white and Asian Pacific Islander among high income group. 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.