Goal: To Determine Whether a Person Earns over $50k using Classification Methods

Description of “Adult” dataset:

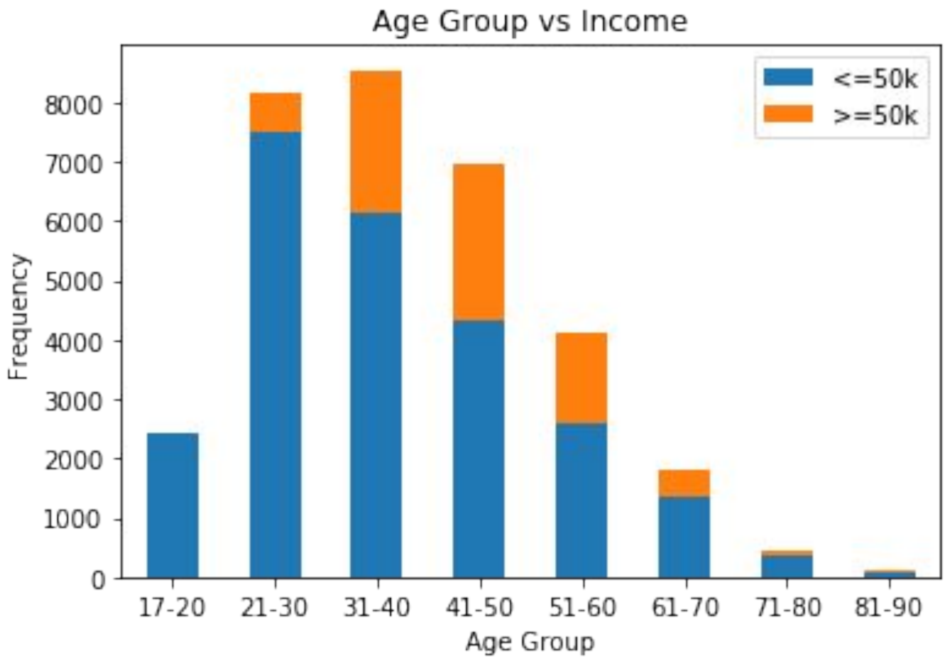
The “**1994 United States Census Income**” dataset, also known as the “**Adult**” dataset contains 14 attributes including one class attribute and 32561 instances.

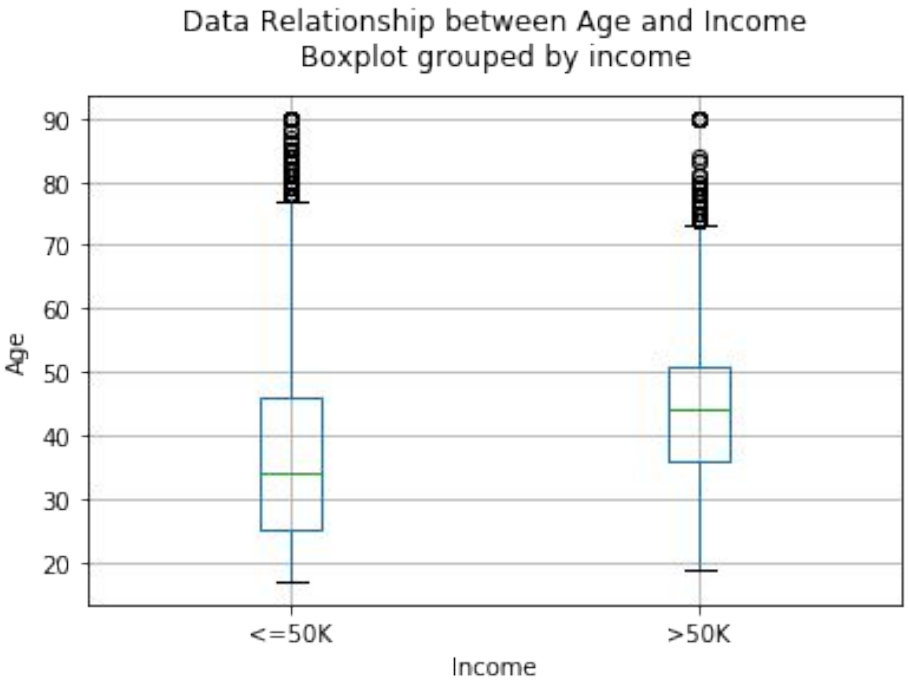
- **Attributes such as:** age, occupation, workclass, education, working hours per week, etc...
- **1 class attribute:** Income of a person (>$50k or <=$50k)
- Attributes are used to predict the class attribute (income).

Data Preparation:

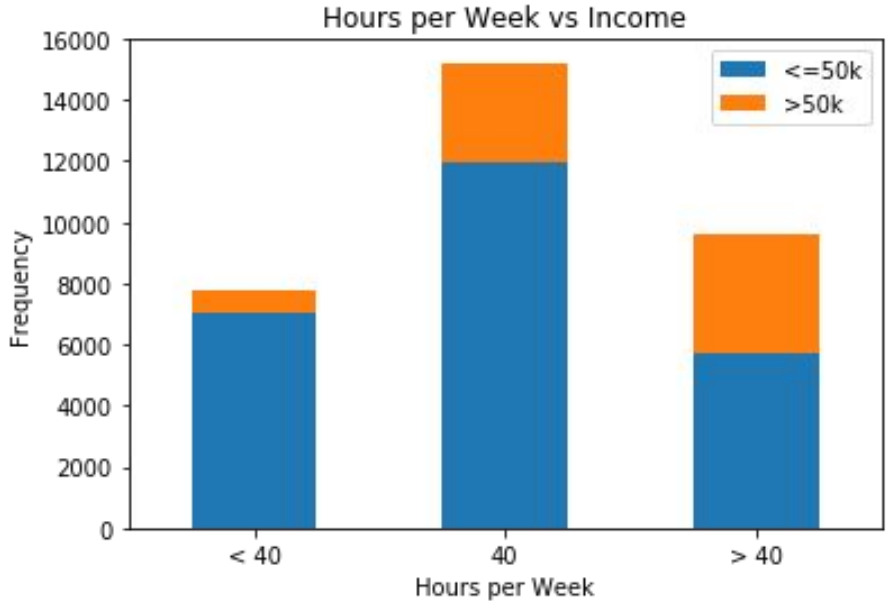
- Replaced missing values represented by “?” in ‘workclass’, ‘occupation’ and ‘native-country’ with their most frequent value respectively.

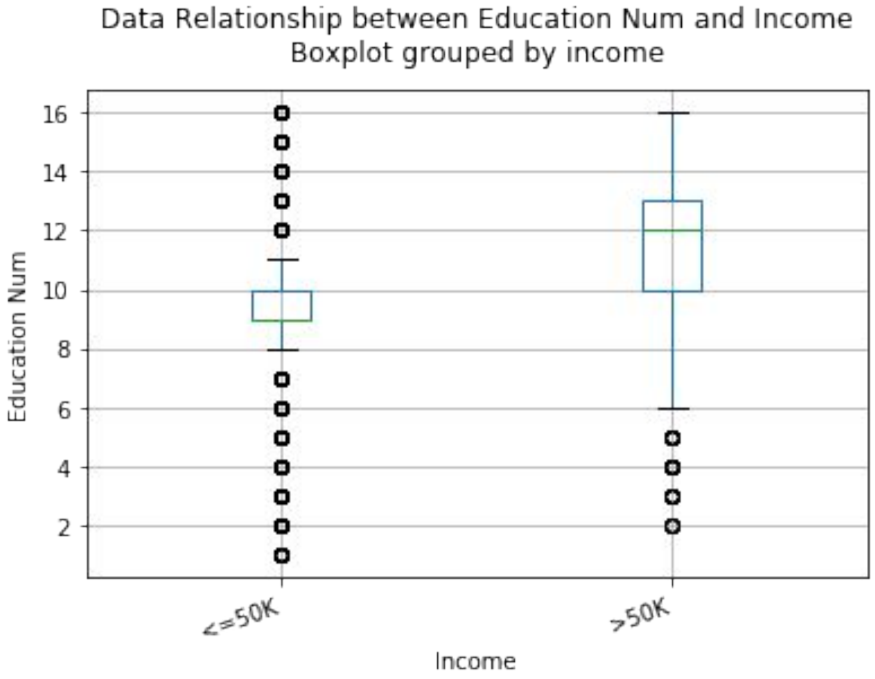
Hypotheses:





Hypotheses (continued…):





Data Modelling Analysis:

Classification Methods	f1-score Weighted Average Score	Error Rate (10-Cross-Validation)
K-Nearest-Neighbour Classifier (*k=5, weights='distance', p=1)*	0.76	0.2192
DecisionTree Classifier (*criterion="entropy", max_depth=12, min_samples_split= 6)*	**0.85**	**0.1439**

Task 1: 50% training / 50% testing

Classification Methods	f1-score Weighted Average Score	Error Rate (10-Cross-Validation)
K-Nearest-Neighbour Classifier (*k=5, weights='distance', p=1)*	0.76	0.2192
DecisionTree Classifier (*criterion="entropy", max_depth=12, min_samples_split= 6)*	**0.85**	**0.144**

Task 2: 60% training / 40% testing

Data Modelling Analysis (continued…):

Classification Methods	f1-score Weighted Average Score	Error Rate (10-Cross-Validation)
K-Nearest-Neighbour Classifier (*k=5, weights='distance', p=1)*	0.76	0.2192
DecisionTree Classifier (*criterion="entropy", max_depth=12, min_samples_split=6)*	**0.85**	**0.144**

Task 3: 80% training / 20% testing

Recommendations:

- Highest level of education achieved by one affects income, therefore we recommend achieving high level of education.
- Get a good job and do well, you will definitely get a good pay.
- For data modelling:
  a. We recommend using *DecisionTree Classifier* with the following parameters:
     i. *criterion="entropy", max_depth=12, min_samples_split=6*
  b. *DecisionTree Classifier* provides a **higher *f1-score Weighted Average Score*** and a **lower *Error Rate*** than using *K-Nearest-Neighbour Classifier* as seen in the tables.