In [1]:
```
import pandas as pd
```

Read census data from data directory

In [2]:
```
df = pd.read_csv('./data/census.csv', header=0, index_col=False)
```

Have a look at the data

Get information about the dataframe's dtypes, missing values, column names

In [3]:
```
df.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    fnlgt           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   salary          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

Result: No missing values; some numerical and non-numerical data (columns); column names contain white spaces

Remove white spaces from the column names

In [4]:
```
# column names before stripping white spaces
df.columns
```

Out[4]:
```
Index(['age', ' workclass', ' fnlgt', ' education', ' education-num',
       ' marital-status', ' occupation', ' relationship', ' race', ' sex',
       ' capital-gain', ' capital-loss', ' hours-per-week', ' native-country',
       ' salary'],
      dtype='object')
```

In [5]:
```
df.columns = [col.strip() for col in df.columns]
```

In [6]:
```
# column names after stripping white spaces
df.columns
```

Out[6]:
```
Index(['age', 'workclass', 'fnlgt', 'education', 'education-num',
       'marital-status', 'occupation', 'relationship', 'race', 'sex',
       'capital-gain', 'capital-loss', 'hours-per-week', 'native-country',
       'salary'],
      dtype='object')
```

Check if there are string data in the table containing white spacse and if yes, remove them.

In [7]:
```
df.select_dtypes(include = 'object').apply(lambda s: (s.str.startswith(' ') | s.str
```

Out[7]:
```
workclass         True
education         True
marital-status    True
occupation        True
relationship      True
race              True
sex               True
native-country    True
salary            True
dtype: bool
```

Result: There are leading and trailing white spaces in every column containing string values.

In [8]:
```
# Remove white spaces
df[df.select_dtypes(include='object').columns] = df.select_dtypes(include = 'object
```

In [9]:
```
# Check result
df.select_dtypes(include = 'object').apply(lambda s: (s.str.startswith(' ') | s.str
```

Out[9]:
```
workclass         False
education         False
marital-status    False
occupation        False
relationship      False
race              False
sex               False
native-country    False
salary            False
dtype: bool
```

Store cleaned data

In [10]:
```
# We store the cleaned data including an index column as these values are taken
# as reference to slice data (see model.py)
df.to_csv('./data/census_cleaned.csv', index=True, index_label='index')
```