Data Science - Homework 2

- Motivation

Because the dataset that I select for final project is in the process of sorting and collecting, I will use adult dataset from UCI as the dataset for this assignment, it also known as "Census Income" dataset. Therefore, my prediction task is to determine whether a person makes over 50K a year.

- Use Python to analyze the dataset that I select, and explain the result I obtain (IDE: jupyter)

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	gender	capital_gain	capital_loss	hours_per_week
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40

First, after importing the dataset, I found that it has 32561 records and 15 attributes in training set, and then I check its details, as can be seen from the figure on the right that there are missing values in three columns, including workclass, occupation and native_country.

Second, I summarized the class distribution, where about 76% of the population has income <= 50k and only 24% > 50k, so there may be a data imbalance problem.

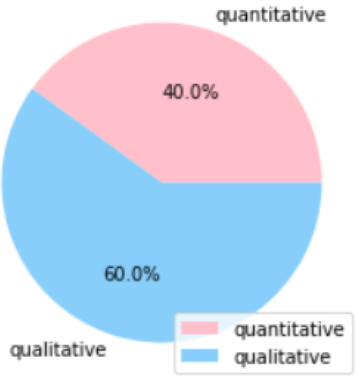
```
df_adult_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
age               32561 non-null int64
workclass         30725 non-null object
fnlwgt            32561 non-null int64
education         32561 non-null object
education_num     32561 non-null int64
marital_status    32561 non-null object
occupation        30718 non-null object
relationship      32561 non-null object
race              32561 non-null object
gender            32561 non-null object
capital_gain      32561 non-null int64
capital_loss      32561 non-null int64
hours_per_week    32561 non-null int64
native_country    31978 non-null object
income_bracket    32561 non-null object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

```
target = df_adult_train.values[:,-1]
counter = Counter(target)
for k,v in counter.items():
    per = v / len(target) * 100
    print('Class: %s, Count=%d, Percentage=%.3f%%' % (k, v, per))
```

```
Class: >50K, Count=7841, Percentage=24.081%
Class: <=50K, Count=24720, Percentage=75.919%
```

Third, I divided training set into quantitative and qualitative. I found that the two columns education and education_num have the same meaning, maybe one of them will be removed when processing in the future.

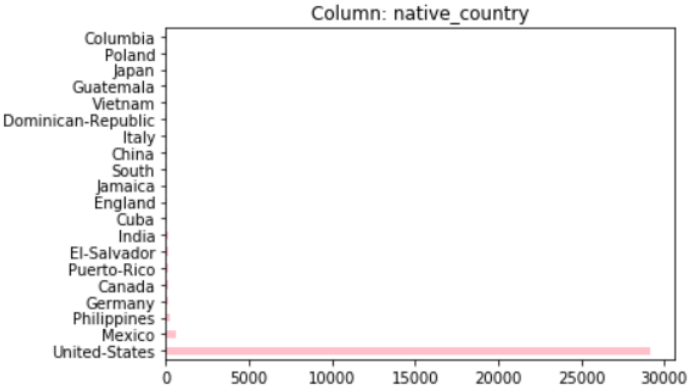


After checking the distributions of qualitative data, I observed quite large number of the native_country=United-States values. This variable can be either ignore or replaced. With binary variable like from_united_states that will be equal to True if person from united states, and False otherwise.

```
for colname in qualitative:
    plt.title('Column: ' + colname)

    (df_adult_train[colname]
         .value_counts()
         .head(20)
         .plot(kind='barh', color=colors[0]))

    plt.show()
```



Finally, my task is to predict the income feature which makes income as dependent variable. If any of the remaining 14 features are affecting the target feature, then those features are known as independent variables.

I ran a for loop over all the columns which gets the count of unique values. We can see that some columns have a lot of distinct values like fnlgwt attribute which has around 20000+ values, it may be noisy data for my model.

```
for c in df_adult_train.columns:
    print ("---- %s ---" % c)
    print (df_adult_train[c].value_counts())
```

```
---- fnlwgt ---
164190    13
203488    13
123011    13
113364    12
121124    12
          ..
284211     1
312881     1
177711     1
179758     1
229376     1
Name: fnlwgt, Length: 21648, dtype: int64
```

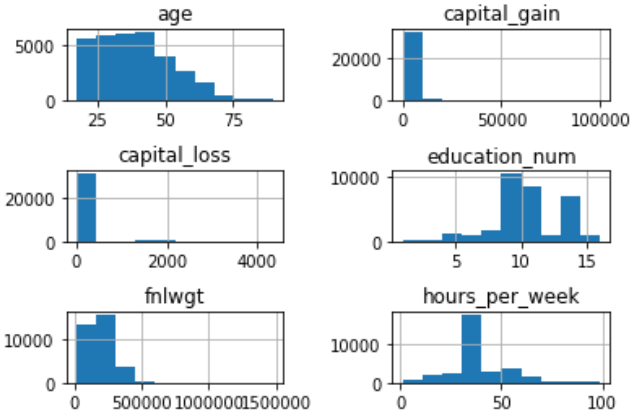
- Discuss possible problems I plan to investigate for future studies

After analyzing this dataset, the first problem I will encounter is missing values. There are two methods to deal with it, one is to drop the rows with missing values, and the other is to impute the missing values with mean or the most frequent value.

Then, the next is the data imbalance problem, because this dataset that contains 76% majority class and 24% minority class. Obviously, the algorithm may fail in its ability to classify the minority class. But is this figure of 24% reflect the real-world problem? If the answer is yes, the dataset is not imbalanced to the extent that it warrants the application of handling of imbalanced.

We can also see that the dataset is a mixture of numerical and categorical or ordinal data types, where the non-numerical columns are represented using strings. These two data type need to be handled with different encoding methods (one-hot encoding or label encoding).

In the numerical data, I found through the histogram that capital_loss and capital_gain have many zero values. I will consider removing them in the future, and it will not affect the prediction results of model.



Moreover, to avoid irrelevant data for my model, I will avoid the less unique or noisy features from our input data. This helps to refine our features to identify the independent variables. Let’s drop the attributes that have noisy data, such as fnlgwt.