DataXpert@UiTM Siri ke-6 (Final Project)

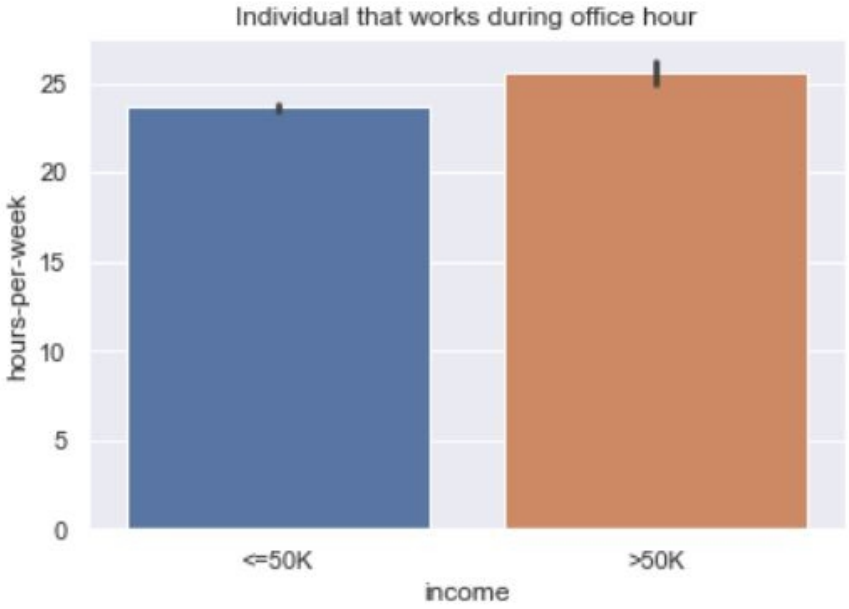
Hexabyte

Predicting Yearly Income

1) **Does working more hours lead to higher income?**

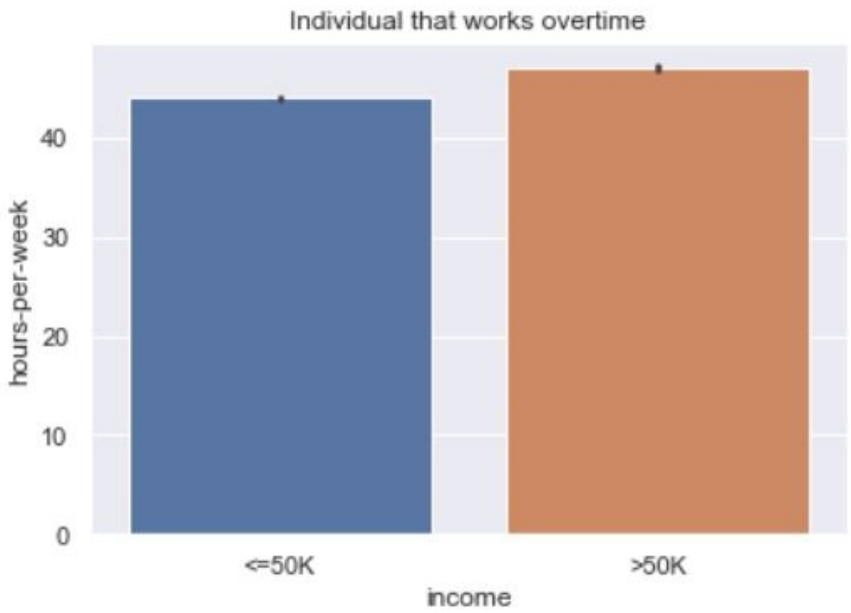
```
Individuals works not overtime and income is more than $50,000: 864
Individuals works not overtime and income is at most $50,000: 9456
Individuals works overtime and income is more than $50,000: 10817
Individuals works overtime and income is at most $50,000: 27653
```

Based on the data given, extract hours-per-week and income based on the condition given that satisfy the question. It is found out that there is a small changes on the data.



Graph 1: Individual that works during office hour

There is a small changes on the graph between the individuals that have low income, >50K, and high income, <=50K. The low income group is smaller than the high income group on the individuals that works during the office hours.



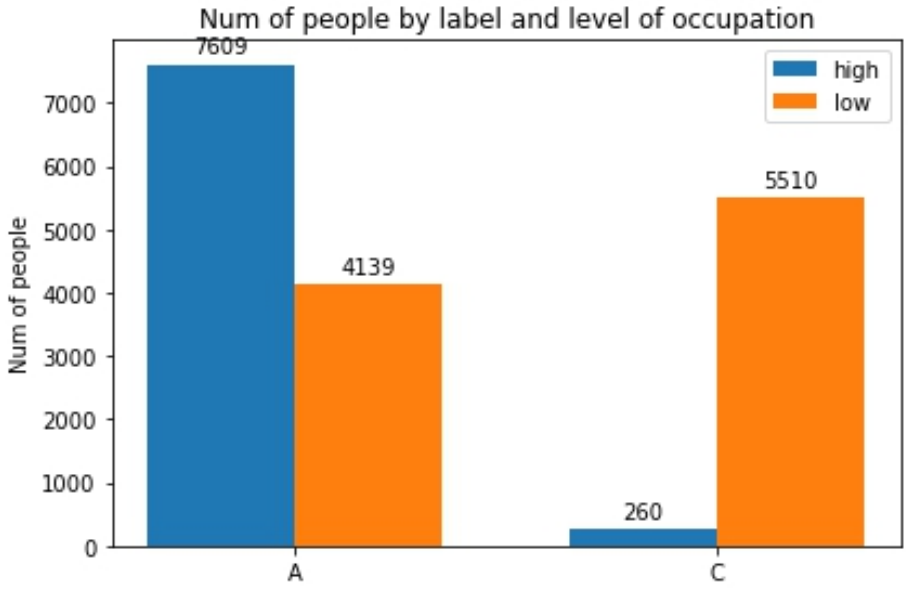
Graph 2: Individual that works overtime

It can be compared from the previous graph. It does not have a big difference between these two group of income. It can be conclude that the working more hours does **not** lead to higher income.

2) **Many people believe that the "A" students will end up working for "C" students. Do you think the statement is correct (based on this dataset)?**

```
A student with high oc: 7609
A student with low oc: 4139
C student with high oc: 260
C student with low oc: 5510
```

These data is extract from the census.csv file and it is found out that 'A' students will **not** end up working for 'C' student. The graph below will shows the details



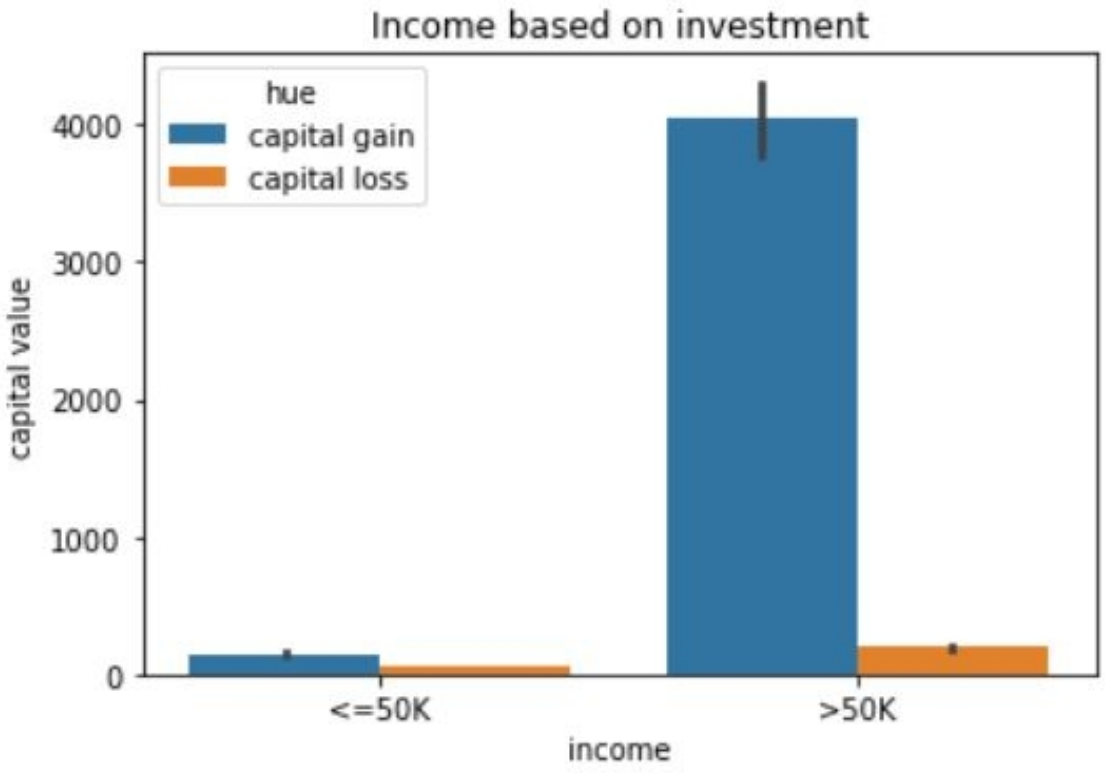
Graph 3: Number of people by label and level of occupation

Most of 'A' student ended up works for a higher occupation than one-thid of them. To contrast the question given, 'C' student has low in high occupation than the rest of them. It can be said that almost every 'C' student works under 'A' student.

3) People with high-income level has a lot of investments?

```
Number of records where individual's income is more than $50,000: 11681
Number of records where individual's income is at most $50,000: 37109
Number of records where individual's capital gain: 4035
Number of records where individual's capital loss: 2282
```

The data is analysis from the fact before transform it into a compare bar graph

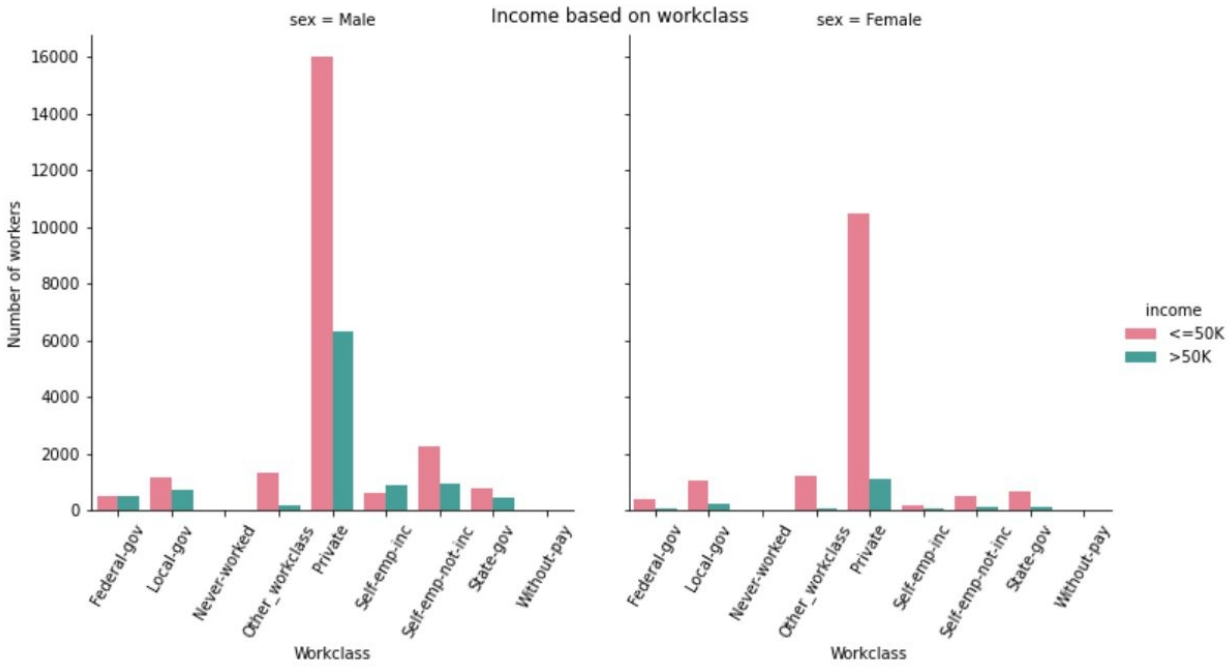


Graph 4: Income based on investment

As shown in the graph 4, high income invests more in the investment and gain profit due to their knowledge in investment. Even their capital loss is small. For the low income, the capital value between gain and loss is not very different rather than the high income. Thus, it is a **true statement** for this question

Develop a machine learning model to predict the yearly income:

Hypothesis:



Graph 5: Income based on work class

To understand more about question 3, a detail analysis is made. Take note that the other_workclass is actually a null value in the data, so do not pay attention too much on it. Most male in high income work in the private sector and second to this, some male workers stay in the self employed not incorporated sector. For the female, most of them works in private sector and local government. As seen in both graph, there is a lot of male than female workers because the male needs to support the family.

Based on the question given, we used supervised with sckit learn, the KNN method due to classify the income of high income, >50K, and low income, <=50K.

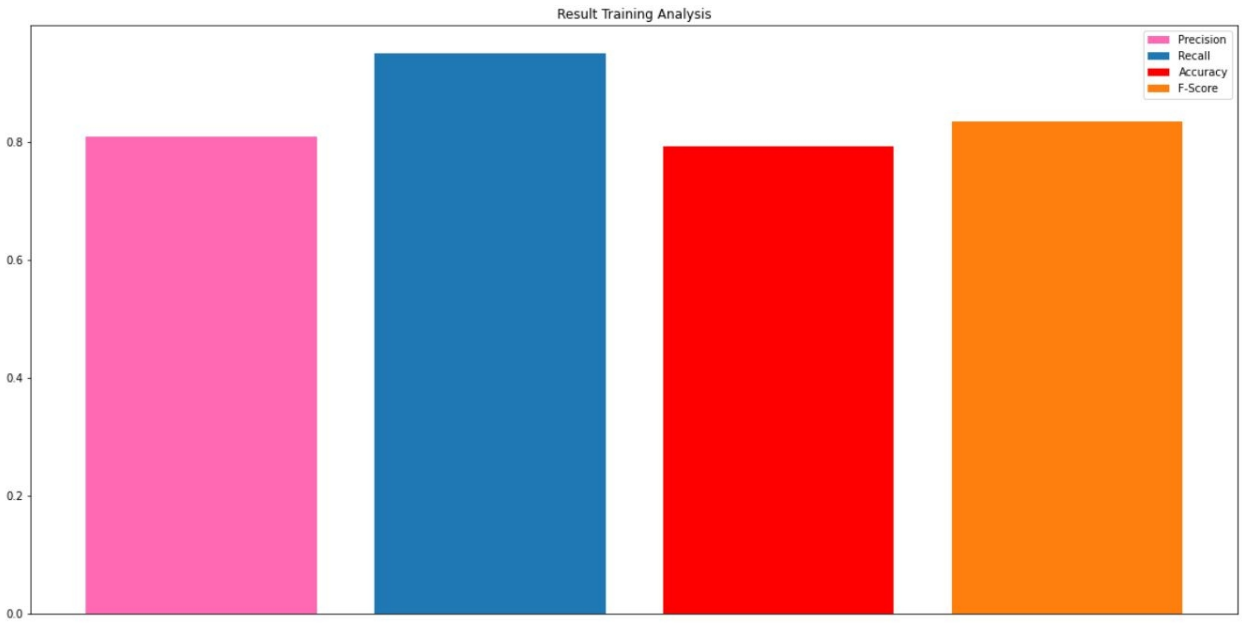
```
KNN: [Accuracy : 0.7916, Precision: 0.8092, Recall: 0.9499, F-score: 0.8339]
```

Accuracy : how much the total observation match with the predicted observation

Precision : how many that is labeled, is actually correctly labeled

Recall : sensitivity, of all, the observation made, how many did we labeled

F-score : to observe uneven class distribution



Graph 6: Result Training Analysis

The Recall is the highest of the data. This is because the column is correctly labeled in the table data. The accuracy is 0.7916 which is quite high for this data.