Project Milestone for CSE578

Yu Tung Lin

Arizona State University
ylin364@asu.edu

Goals and a business objective

The main objectives of this project are to predict income levels, identify factors that influence income levels, and improve data visualization techniques. This dataset may be useful for businesses and organizations that want to target their marketing efforts towards individuals with higher income levels or for government agencies that need to identify and assist individuals who are below the poverty line.
The purpose of the project report is to use data visualization tools to demonstrate an understanding of how to solve the issue of forecasting income levels using a dataset with several crucial variables. That is, this project seek to utilize the application to forecast an individual's income by considering various input parameters. The chosen approach is to group and analyze the factors that may play a crucial role in predicting income and use this information to create marketing profiles for individuals. By forecasting an individual's income based on various input parameters, businesses can customize their marketing strategies when dealing with these individuals.

Assumptions

For some of the features provided, I have some assumptions about the income:

- Age:
  I assume that older individuals have more work experience and, therefore, may have higher salaries compared to young people. However, for the people too old, they may be too old to work, therefore, they will have low salaries.

- Workclass:
  Individuals who work in certain workclass, such as government or finance, may have higher salaries than those who work in other sectors, such as retail or hospitality.

- Education:
  Individuals with higher levels of education, such as a bachelor's or master's degree, may be more likely to have high-paying jobs. That is because, they may have more knowledge than others and people will pay more for higher education workers.

- Education-num:
  This feature is related to education, and it measures the number of years of education an individual has completed. It is likely that individuals with more years of education may have higher salaries just like the above feature.

- Marital-status:
  It is possible that individuals who are single may have higher salaries than those who are married or in a long-term relationship because single people might put more emphasis on work.

- Occupation:
  Different occupations have different salary ranges. For example, doctors and lawyers typically earn higher salaries than teachers or retail workers.

- Relationship:
  This feature indicates the individual's relationship status. It may be that individuals who are the head of the household or who have children may have higher salaries because they need more money to take care of the family.

- Race:
  If there are biases and discrimination in the workplace based on an individual's race, white people may have higher salary.

- Sex:
  Similarly, gender discrimination can occur in the workplace, with women typically earning less than men.

- Capital-gain and capital-loss:
  It is possible that individuals with higher capital gains may have higher salaries because they might be smarter.

- Hours-per-week:
  Individuals who work longer hours may have higher salaries because they care more about their work.

- Native-country:
  It is possible that individuals who come from other countries may have lower salaries because of the language or education difference.

User Stories with Visualizations

- I would like to determine the correlations between the features in my dataset. Specifically, for the "sex" feature, I have assigned a value of 1 for male and 0 for female. Similarly, for the "income" feature, a value of 1 represents an income greater than 50k, while a value of 0 represents an income less than or equal to 50k.

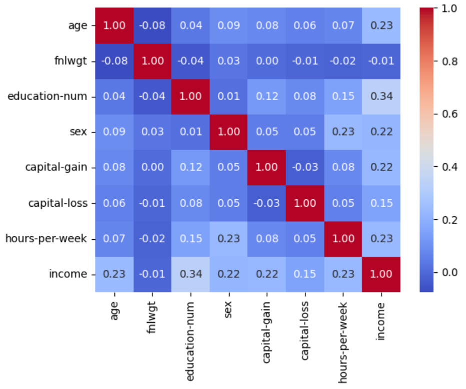


Figure 1: heat map

The highest value in the heat map is 0.34, which is the correlation of income and the education-num.

- I want to know the relation between the age and income.

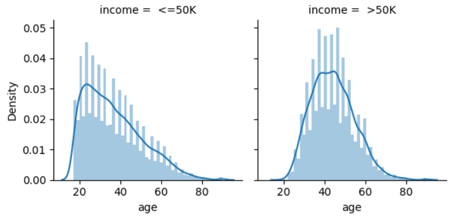


Figure 2: relation between the age and income

The figure shows that younger individuals are more likely to have an income of <=50k. Additionally, individuals around the age of 40 tend to have a higher income.

- I want to know the relation between the age, sex, and income.

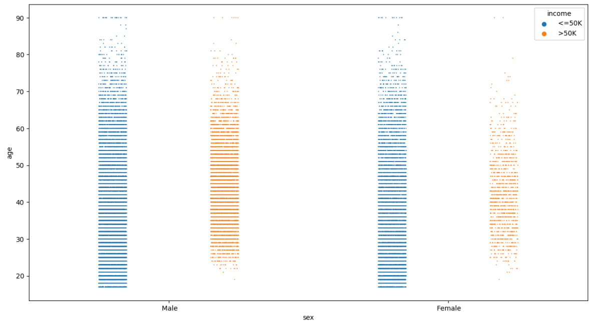


Figure 3: relation between the age, sex, and income

Females have a smaller proportion of individuals with an income greater than 50k. However, for both genders, the age range of 30-60 tends to have a higher proportion of individuals with a higher salary.

- I want to know whether marital status is an impact on income.
  There are many individuals who are not married, and they are more likely to have a lower income. Conversely, individuals in the "married-civ-spouse" category are more likely to have a higher income. As a result, I plan to

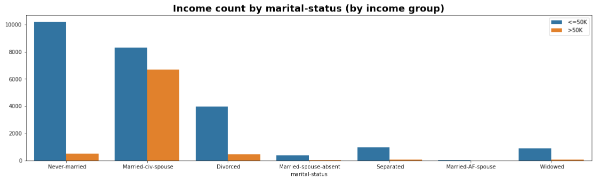


Figure 4: relation between marital status and income

add a new feature indicating whether an individual has a spouse.

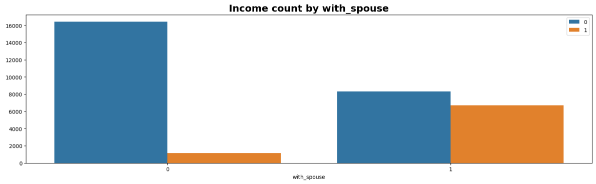


Figure 5: relation between with-spouse and income

The figure shows that people who has spouse has higher income.

- I want to know whether the relationship in a family is a crucial feature for predicting income.

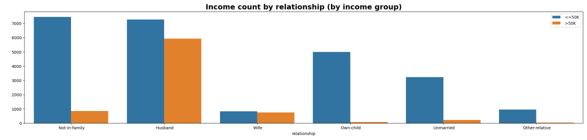


Figure 6: relation between relationship and income

I want to add more categories: mother, father, unmarried male, and unmarried female, since I have the data of their sex.

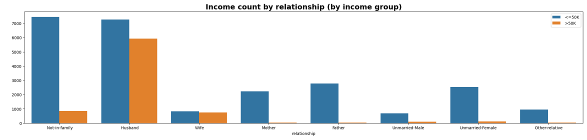


Figure 7: relation between relationship and income

From the above 2 figures, we can see husband and wife has high probability with income > 50k. As a result, I want to add a feature which is whether people is alone.

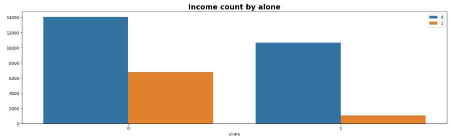


Figure 8: relation between alone and income

- I want to know the relationship with capital gain or loss and income.

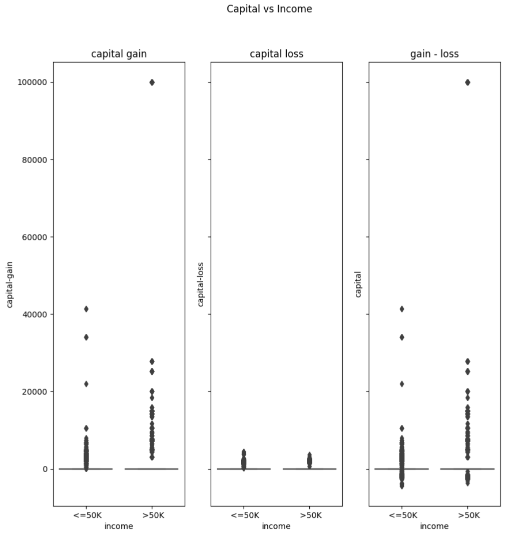


Figure 9: relation between capital and income

I have used a box plot to visualize the data. However, due to the majority of the data being zero, the box in the plot appears as a line at zero. Therefore, I am considering creating a new feature to indicate whether the data point is equal to zero or not.

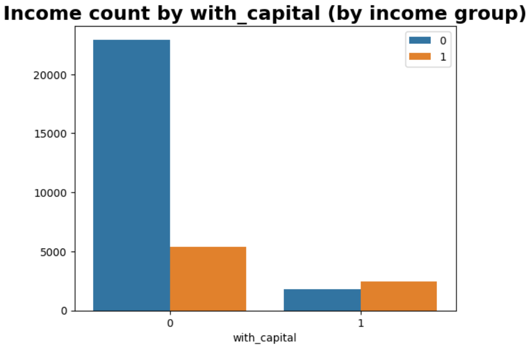


Figure 10: relation between capital and income

- I want to know the relationship between hours-per-week and income.

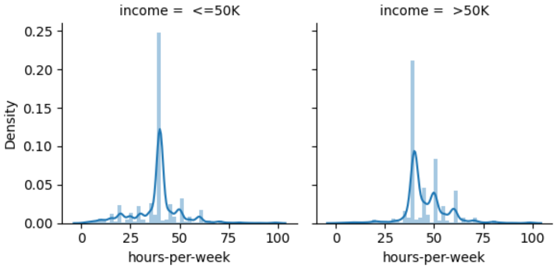


Figure 11: relation between hours-per-week and income

The hours-per-week is a continuous data, I would like to classify it to lesser, normal and more hours.

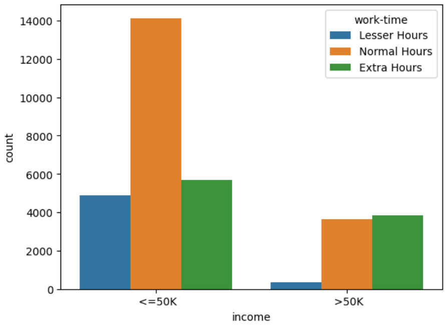
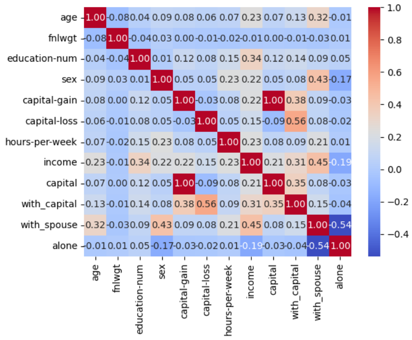


Figure 12: relation between work-time and income

In this figure, it is less likely to have high salary if working less time.

- After adding several features, I'll like to see the correlations between them.



	age	fnlwgt	education-num	sex	capital-gain	capital-loss	hours-per-week	income	capital	with_capital	with_spouse	alone
age	1.00	-0.08	0.04	0.09	0.08	0.06	0.07	0.23	0.07	0.13	0.32	-0.01
fnlwgt	-0.08	1.00	-0.04	0.03	0.00	-0.01	-0.02	-0.01	0.00	-0.01	-0.03	0.01
education-num	0.04	-0.04	1.00	0.01	0.12	0.08	0.15	0.34	0.12	0.14	0.09	0.05
sex	0.09	0.03	0.01	1.00	0.05	0.05	0.23	0.22	0.05	0.08	0.43	-0.17
capital-gain	0.08	0.00	0.12	0.05	1.00	-0.03	0.08	0.22	1.00	0.38	0.09	-0.03
capital-loss	0.06	-0.01	0.08	0.05	-0.03	1.00	0.05	0.15	-0.09	0.56	0.08	-0.02
hours-per-week	0.07	-0.02	0.15	0.23	0.08	0.05	1.00	0.23	0.08	0.09	0.21	0.01
income	0.23	-0.01	0.34	0.22	0.22	0.15	0.23	1.00	0.21	0.31	0.45	-0.19
capital	0.07	0.00	0.12	0.05	1.00	-0.09	0.08	0.21	1.00	0.35	0.08	-0.03
with_capital	0.13	-0.01	0.14	0.08	0.38	0.56	0.09	0.31	0.35	1.00	0.15	-0.04
with_spouse	0.32	-0.03	0.09	0.43	0.09	0.08	0.21	0.45	0.08	0.15	1.00	-0.54
alone	-0.01	0.01	0.05	-0.17	-0.03	-0.02	0.01	-0.19	-0.03	-0.04	-0.54	1.00

Figure 13: heat map with more features

The highest correlation with the income has become with spouse. After visualizing the correlation, I want to see pairwise relationships in the dataset.

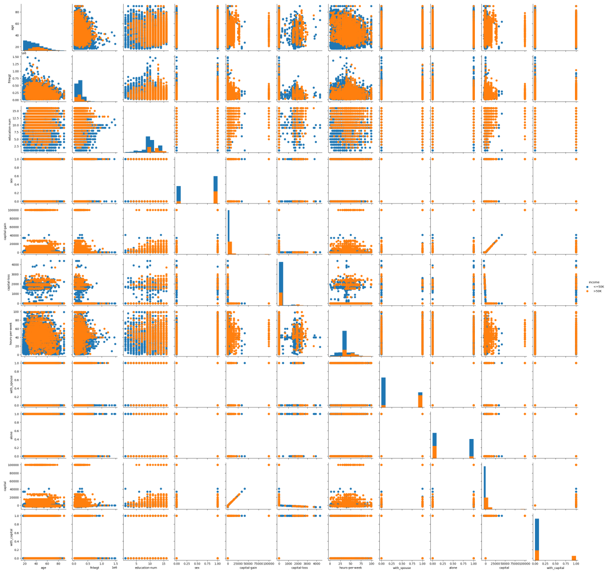


Figure 14: pairwise relationships

With the orange part meaning higher salary, i can get the boundary for each class. For example, people in the education-num higher than 2.5 are more likely to have income >50k. With these boundary, I can categorize feature with continuous values and make the data smaller for training.

- I want to know whether education-num is related to occupation.
  To address the issue of missing data represented as '?' in the occupation column, I have formulated a plan. For instances where occupation = '?' and workclass = never-worked, I will add a new class called "no-work" to the occupation column and include these individuals in that category. However, for all other instances where occupation = '?' and workclass is also equal to '?', I have decided to explore the education-num column for clues. After reviewing the data, I have concluded that I will fill all of the remaining missing data in the workclass column with "private" since this category has the highest frequency. Regarding the occupation column, I have decided to assign in-

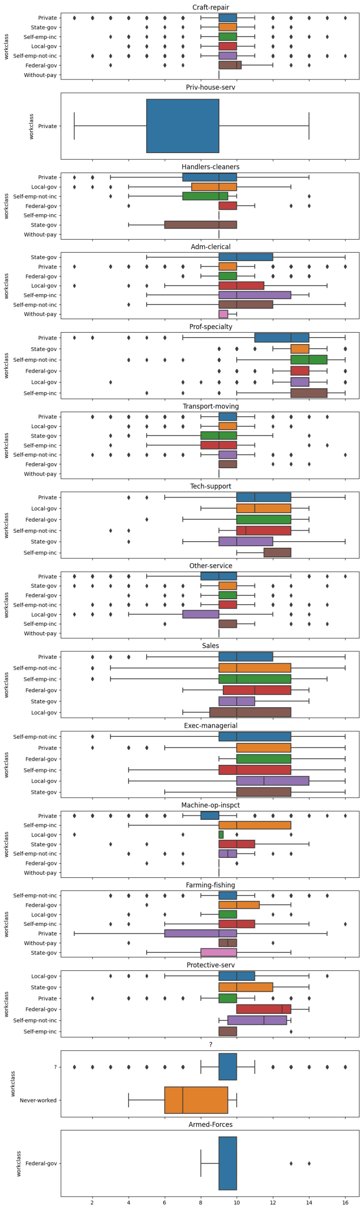


Figure 15: education-num vs occupation vs workclass

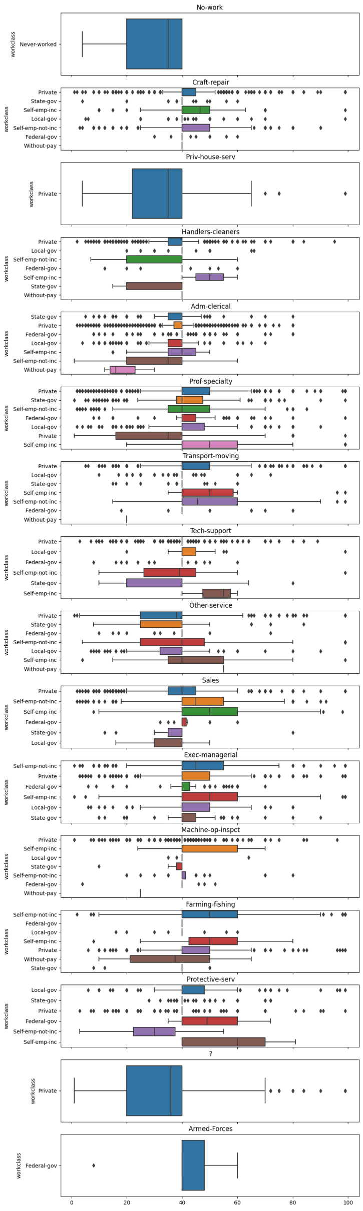


Figure 16: hours per week vs occupation vs workclass

dividuals with an education-num value greater than or equal to 12 to the ”Prof-specialty” category. For the remaining missing values, I will check the work-hours column to make a more informed decision.

The distribution for occupation = ’?’ and workclass = private looks a lot like the distributions for occupations such as priv-house-serv, handlers-cleaners, Adm-clerical, and other-services. Therefore, it is difficult for me to make a decision. However, after looking at the figure that shows hours per week vs occupation vs workclass, I can conclude that it should not be Adm-clerical. Next, I will check the visualization of occupation vs income.

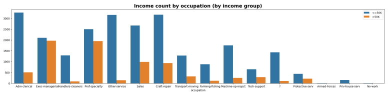


Figure 17: occupation vs income

Since, the '?' category distribution match other-service more, I'll classify the '?' into other-service.

- I want to know whether native-country is related to race.

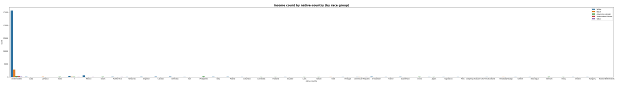


Figure 18: race vs country

As shown in the figure, the majority of people are from the USA, and the USA has almost all races of people. For the '?' in the country, I have decided to choose the race of white, black, and other and assign them to the USA, while the race of Asian-Pac-Islander will be assigned to the Philippines.

- I want to know whether education-num is related to income.

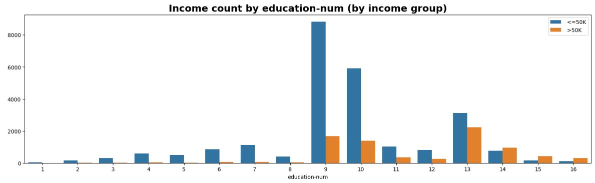


Figure 19: education-num vs income

Since the lower numbers(1-8) have less people in it, I'll group them together; since the higher numbers(15,16) have less people in it, I'll group them together.

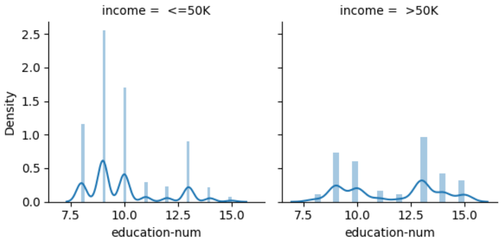
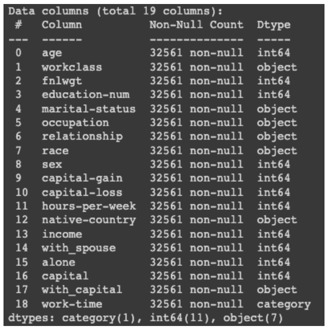


Figure 20: education-num vs income

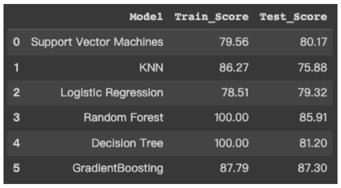
- Before making the model, I want to make sure what features will be the input.



```
Data columns (total 19 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-num   32561 non-null  int64
 4   marital-status  32561 non-null  object
 5   occupation      32561 non-null  object
 6   relationship    32561 non-null  object
 7   race            32561 non-null  object
 8   sex             32561 non-null  int64
 9   capital-gain    32561 non-null  int64
 10  capital-loss    32561 non-null  int64
 11  hours-per-week  32561 non-null  int64
 12  native-country  32561 non-null  object
 13  income          32561 non-null  int64
 14  with_spouse     32561 non-null  int64
 15  alone           32561 non-null  int64
 16  capital         32561 non-null  int64
 17  with_capital    32561 non-null  object
 18  work-time       32561 non-null  category
dtypes: category(1), int64(11), object(7)
```

Figure 21: features

- This is the accuracy of each model I've trained.



	Model	Train_Score	Test_Score
0	Support Vector Machines	79.56	80.17
1	KNN	86.27	75.88
2	Logistic Regression	78.51	79.32
3	Random Forest	100.00	85.91
4	Decision Tree	100.00	81.20
5	GradientBoosting	87.79	87.30

Figure 22: model performance

As the figure shown, the Gradient Boosting Classifier have the best performance of all.

Questions

- What to do with "?" values?
  Initially, I considered dropping these values altogether, but I soon realized that doing so would lead to a loss of valuable data. Instead, I decided to replace the missing values with the mode values in their respective columns. By doing this, I was able to preserve the data and minimize the impact of the missing values on my analysis.

- How is the column "education" and "education-num" related, should i drop one of it?
  After visualizing the data, I discovered that there is a strong correlation between the "education" and "education-num" columns. Specifically, higher values in the "education-num" column correspond to higher levels

of education. As a result, I made the decision to drop the ”education” column, as I believed that the information provided by the ”education-num” column was sufficient for my analysis.

- The missing values are located in the ”workclass”, ”occupation”, and ”native-country” columns. While I am uncertain about how to approach the missing data in the ”workclass” and ”occupation” columns, I do believe that the missing values in the ”native-country” column may be related to race. In order to fill in these values, I plan to examine the relationship between race and country of origin, and use this information to make informed guesses. Additionally, I will explore the ”education-num” data to see if it can provide any additional insights that might help me to fill in the missing data in these columns.

Not Doing

In the future, I aspire to find a better way of handling the ’?’ values in datasets. Rather than relying solely on assumptions, I aim to delve deeper into the data and uncover additional evidence that can provide me with the necessary confidence to fill in the missing values. By utilizing more sophisticated techniques for data imputation, such as machine learning algorithms or statistical models, I hope to arrive at more accurate and reliable results. By understanding the reasons behind missing values and the patterns of their occurrences can also help me design more effective data collection processes that minimize the likelihood of missing data in the future.

Moreover, I recognize that there is still much room for improvement, and I believe that tuning the parameters of the models could lead to significant enhancements in their performance.

While I have already explored various machine learning models and techniques, I realize that there are still many other models that I have not yet tried. In order to expand my skill set and improve the accuracy of my models, I am eager to explore these models further and determine which ones are best suited to the particular dataset and task.

By focusing on the optimization of model parameters, I hope to achieve a higher level of accuracy in my predictions. This will involve experimenting with different hyperparameters, such as learning rates, regularization factors, and activation functions, among others. By carefully tuning these parameters and evaluating the resulting models, I can identify the optimal combination of settings that yield the best results.