Analysis of Annual Income

Tiancheng Liu, Yuhan Zhu, Yan Lyu

May 02, 2023

Content

Introduction

Our project aims to analyze the Census Income dataset, which contains demographic and employment-related information about individuals, with a target variable representing whether an individual's annual income exceeds $50,000. We'll explore the dataset, clean and transform the data, and fit various models to predict whether an individual's annual income exceeds $50,000.

Description of Data

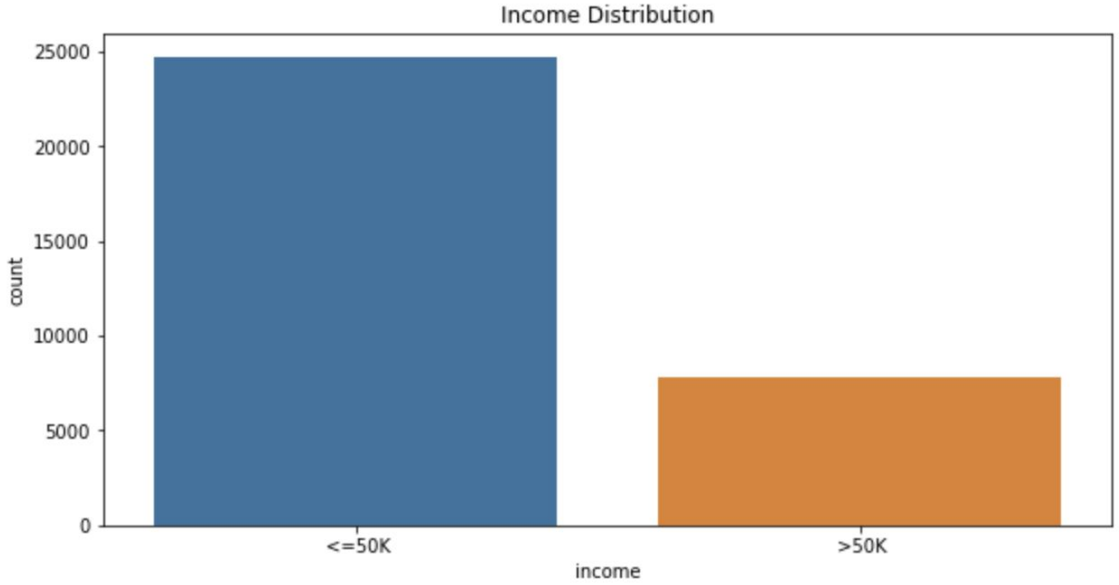
- **What**: The Adult dataset is a collection of demographic and employment related information about individuals from the 1994 Census database, with a target variable indicating whether the individual's income exceeds $50,000 per year.

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	2772466

- **Where**: https://archive.ics.uci.edu/ml/datasets/Adult
- **Attribute Information:** The dataset contains 48842 instances and 14 variables. . Columns will be used include the person's job, region, age, gender, work class, education, etc.

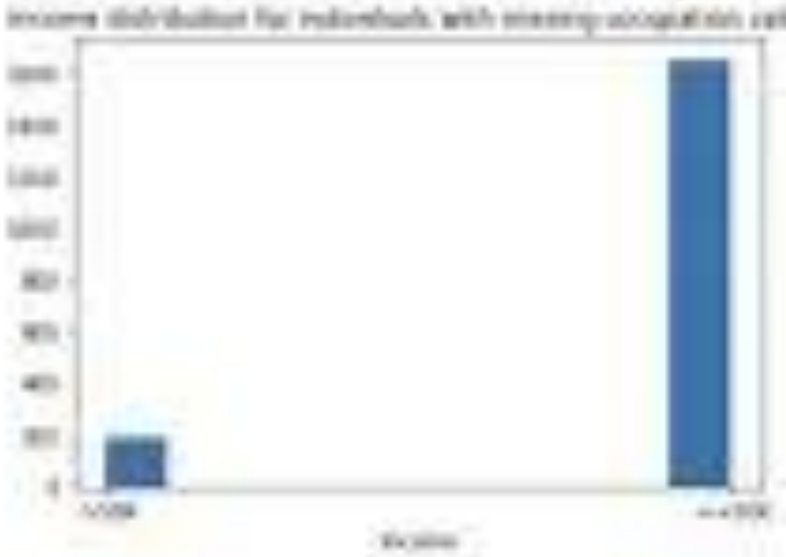
Exploratory Data Analysis

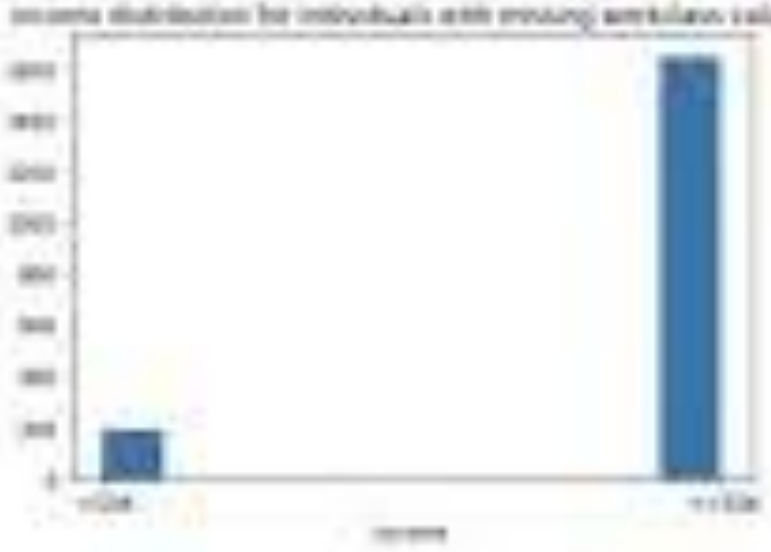
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32560 entries, 0 to 32559
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32560 non-null  int64
 1   workclass       32560 non-null  object
 2   fnlwgt          32560 non-null  int64
 3   education       32560 non-null  object
 4   education_num   32560 non-null  int64
 5   marital_status  32560 non-null  object
 6   occupation      32560 non-null  object
 7   relationship    32560 non-null  object
 8   race            32560 non-null  object
 9   sex             32560 non-null  object
 10  capital_gain    32560 non-null  int64
 11  capital_loss    32560 non-null  int64
 12  hours_per_week  32560 non-null  int64
 13  native_country  32560 non-null  object
 14  income          32560 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```



<=50K	24719	75.92%
>50K	7841	24.08%

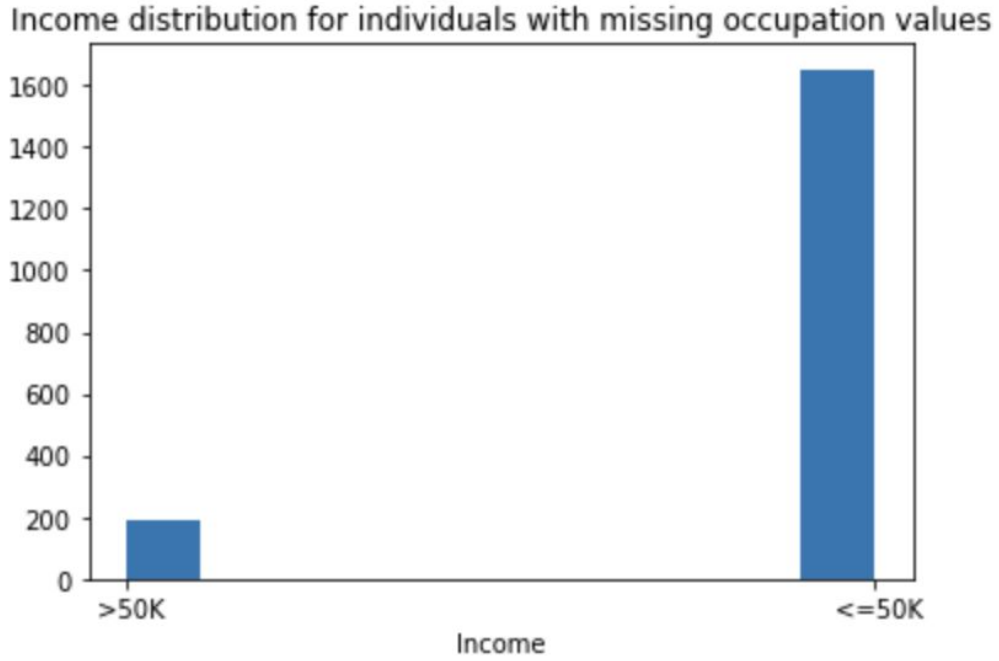
	↑ Count	Percentage
occupation	164000	5.660
workclass	163800	5.630
native_country	50800	1.761
age	0	0.000
fnlwgt	0	0.000
education	0	0.000
education_num	0	0.000
marital_status	0	0.000
relationship	0	0.000
race	0	0.000
sex	0	0.000
capital_gain	0	0.000
capital_loss	0	0.000
hours_per_week	0	0.000
income	0	0.000

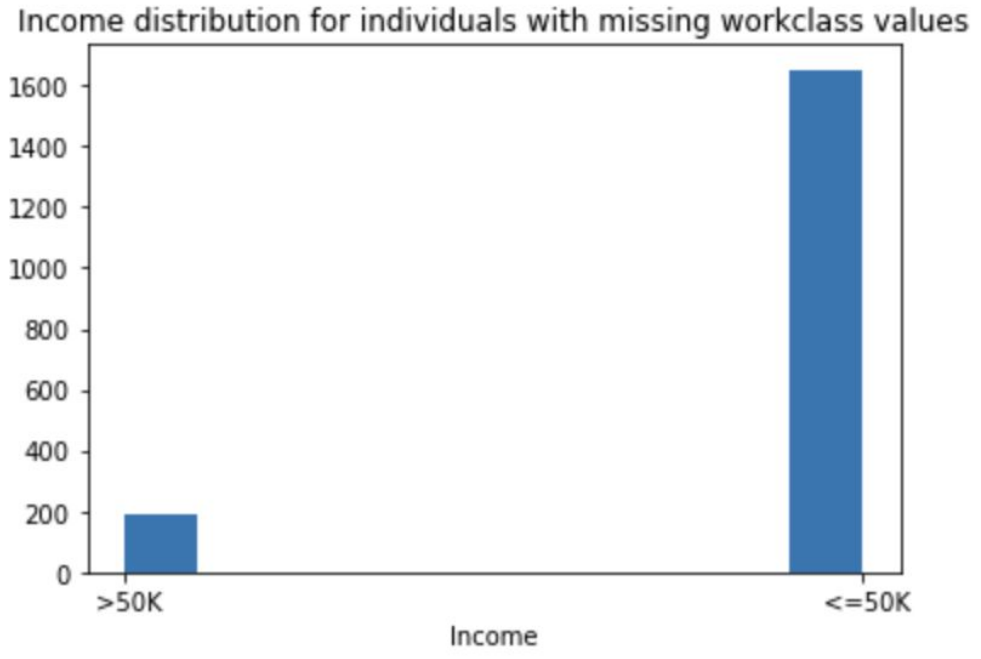




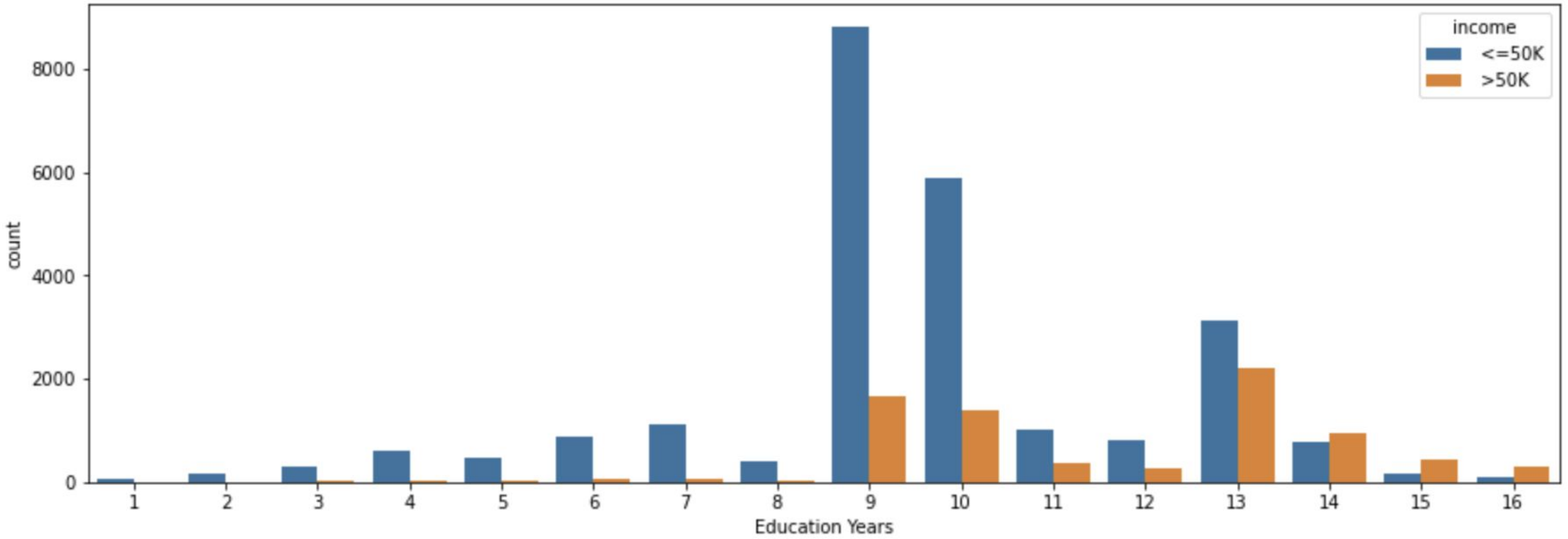
Missing values

	?_Count	Percentage
occupation	184300	5.660
workclass	183600	5.639
native_country	58300	1.791
age	0	0.000
fnlwgt	0	0.000
education	0	0.000
education_num	0	0.000
marital_status	0	0.000
relationship	0	0.000
race	0	0.000
sex	0	0.000
capital_gain	0	0.000
capital_loss	0	0.000
hours_per_week	0	0.000
income	0	0.000



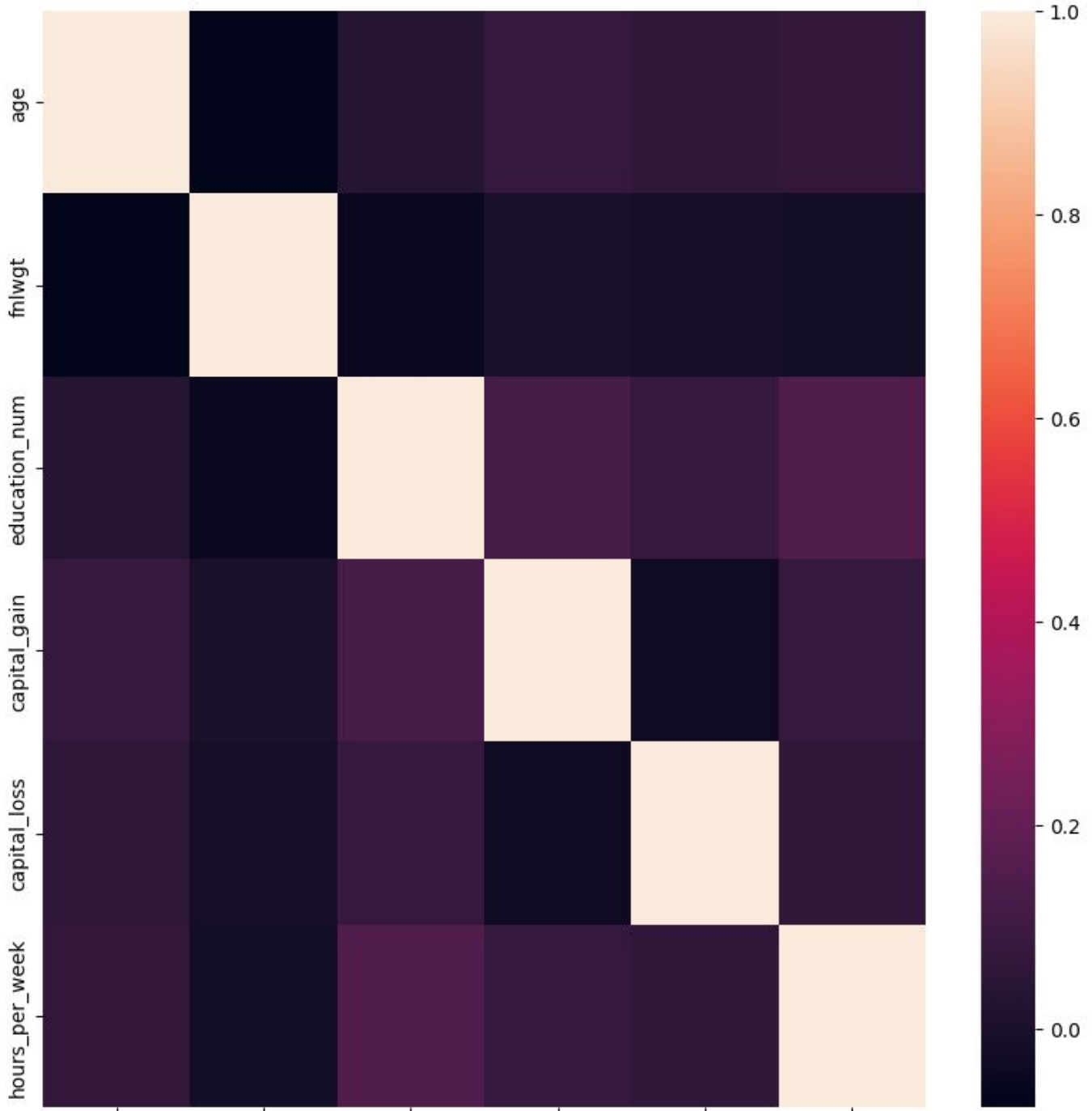


Income by Education Level



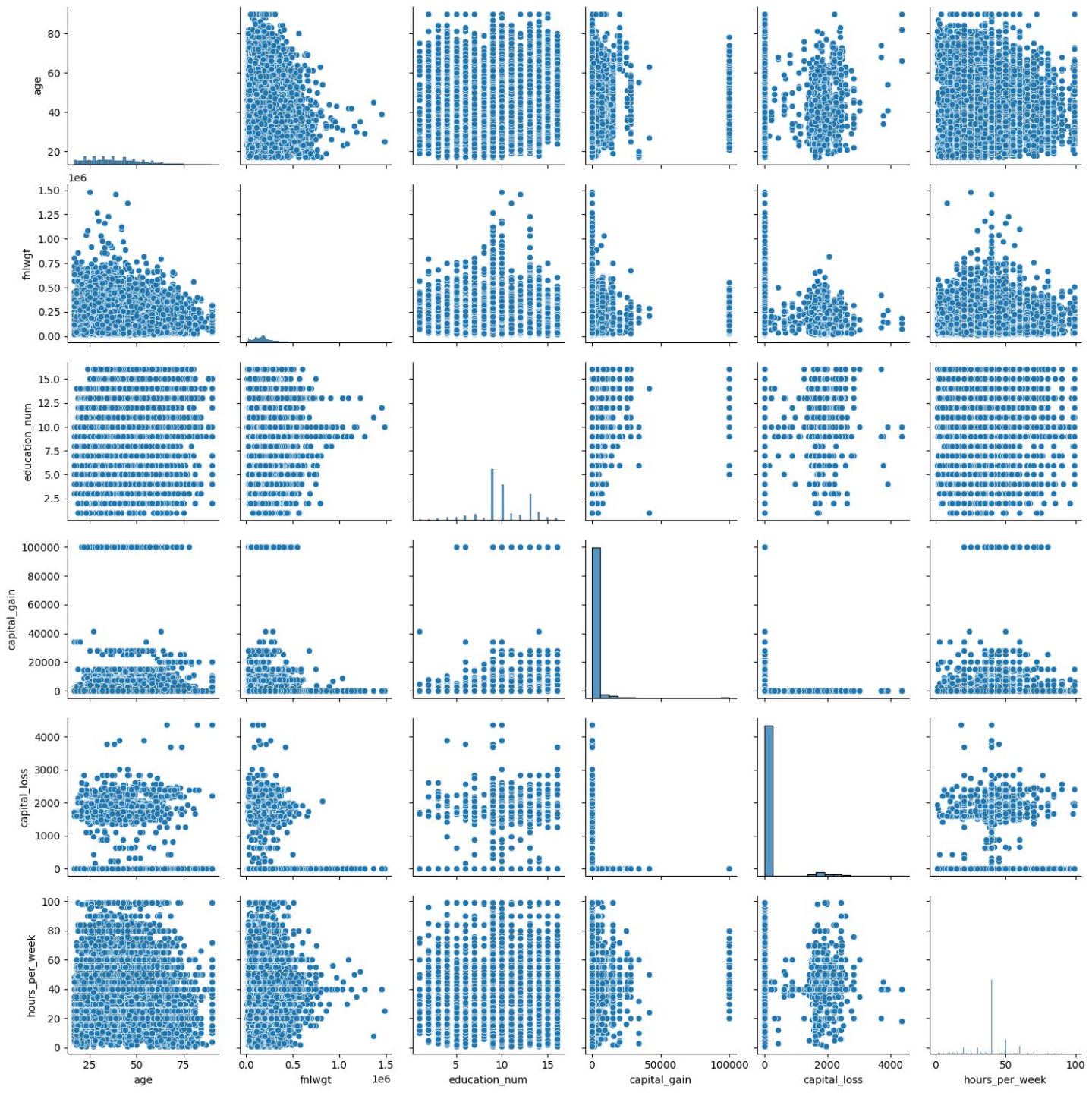
Correlation

- The correlation heatmap suggests that the numerical variables are little correlated to each other, means we can use them all in model fitting without worrying about multicollinearity.



Pairplot on the variables

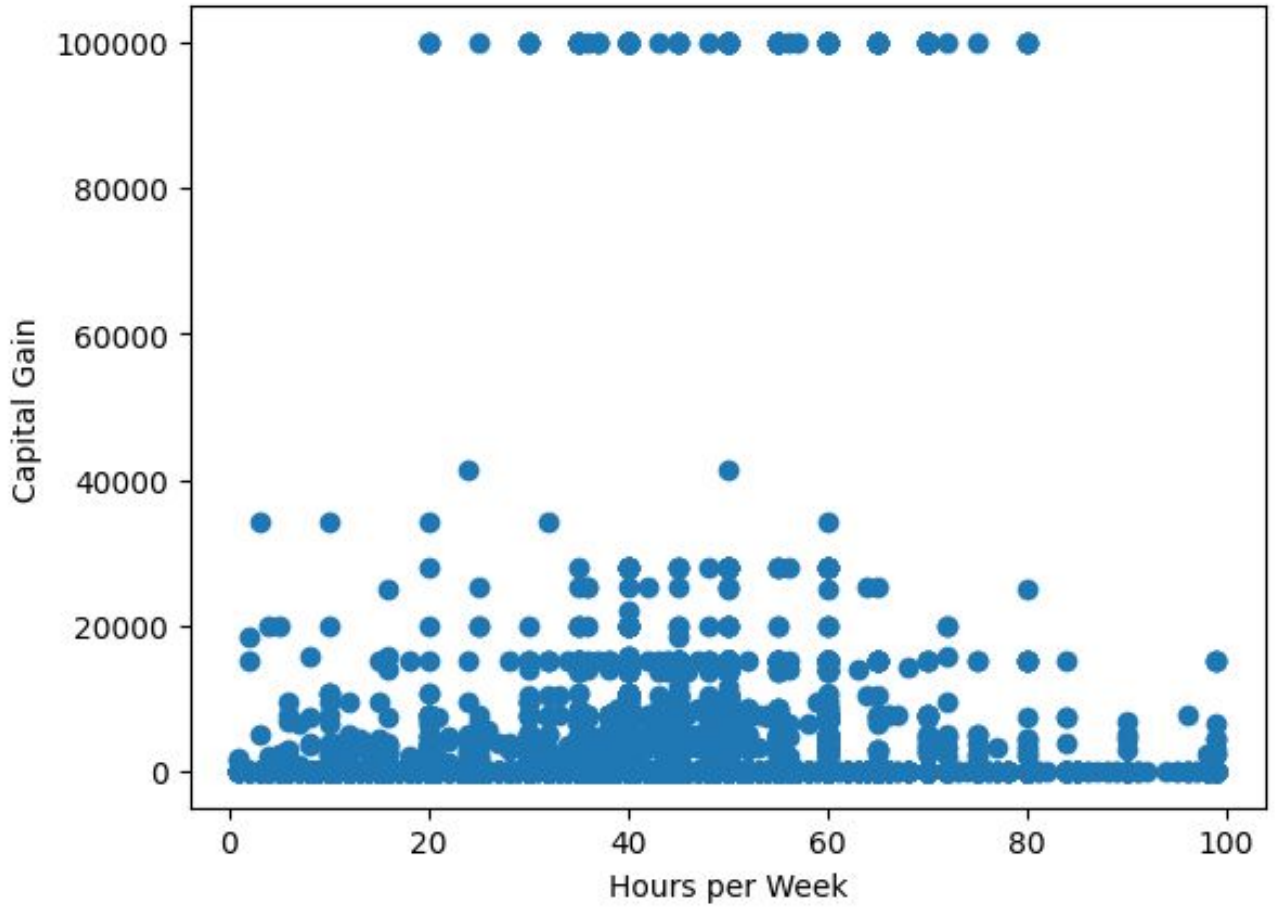
No clear pattern seen.



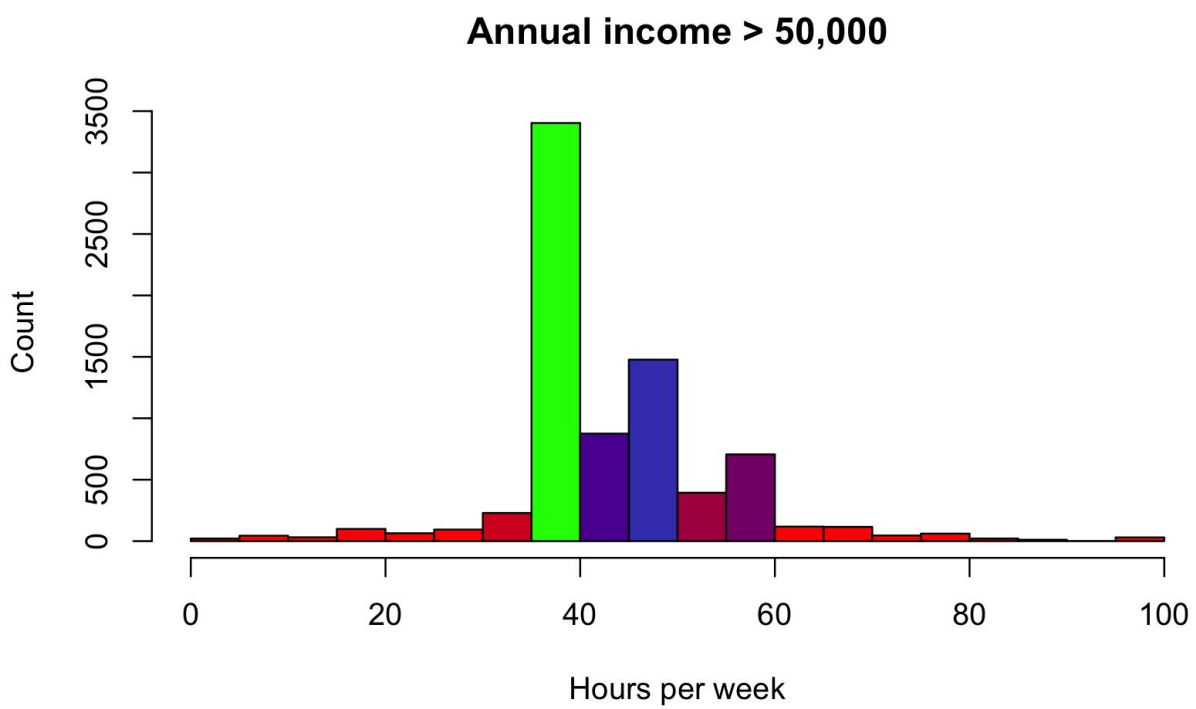
Some findings

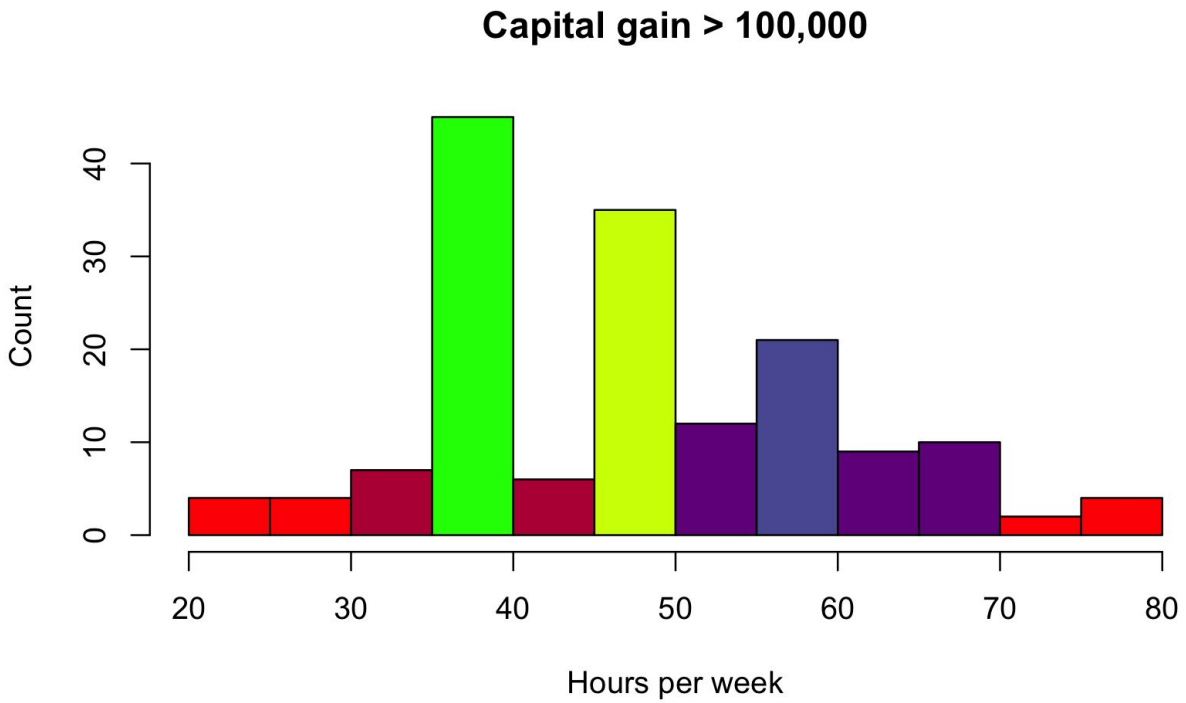
Capital gain/Hours per week

- Normal distribution
- Slightly right skewed



Hong long should you work per week?





Modeling

Model selection:

- Logistic regression(Baseline model)
- Random forest
- XGBoost
- KNN
- K-Means
- Decision Tree
- ...

Baseline model(Logistic regression)

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7000 1722
         1  414  633

               Accuracy : 0.7813
                 95% CI : (0.773, 0.7895)
    No Information Rate : 0.7589
    P-Value [Acc > NIR] : 8.776e-08
```

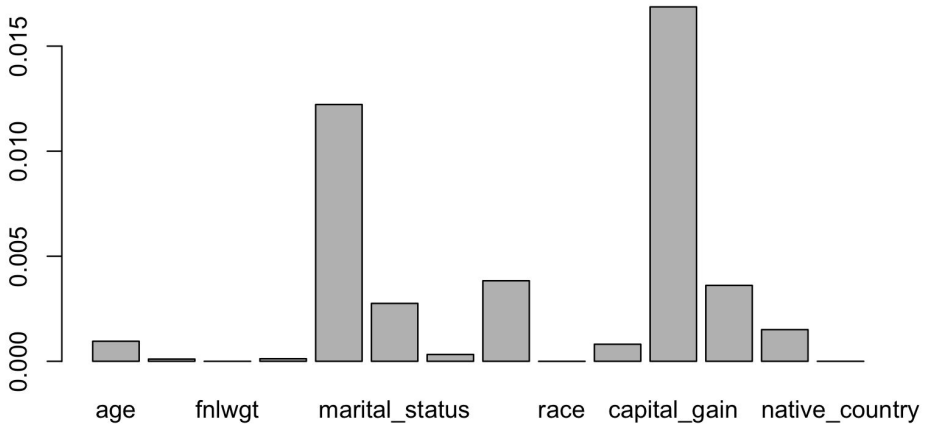
Logistic Regression with aggregation

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4415  993
         1  213  339

               Accuracy : 0.7977
                 95% CI : (0.7872, 0.8078)
    No Information Rate : 0.7765
    P-Value [Acc > NIR] : 3.988e-05
```

Random Forest

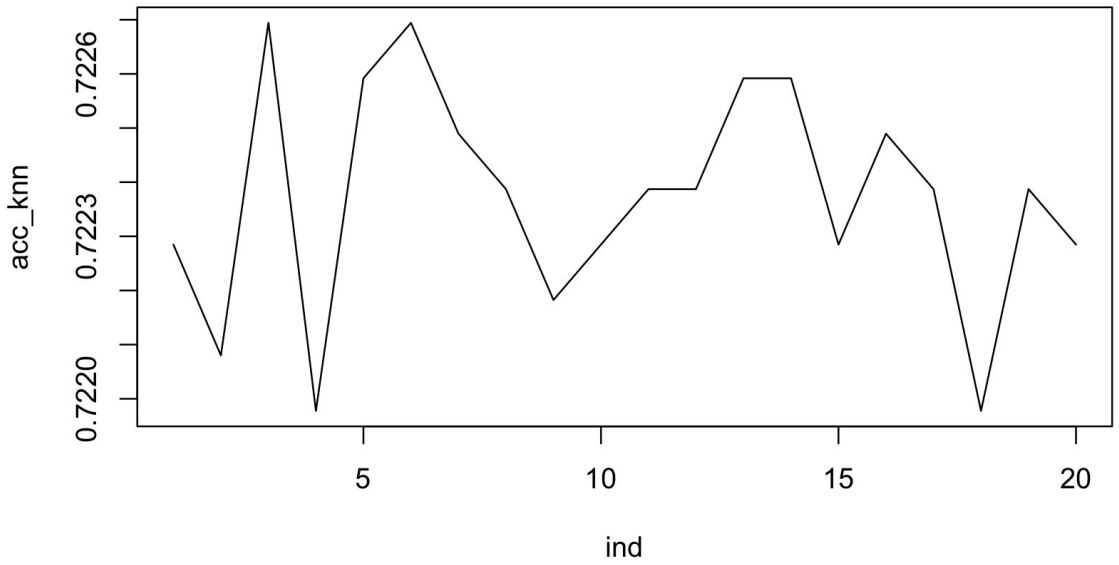



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7399 2370
         1    0    0

               Accuracy : 0.7574
                 95% CI : (0.7488, 0.7659)
    No Information Rate : 0.7574
    P-Value [Acc > NIR] : 0.5055
```

KNN




```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6578 1891
         1  821  479

               Accuracy : 0.7224
                 95% CI : (0.7134, 0.7313)
    No Information Rate : 0.7574
    P-Value [Acc > NIR] : 1
```

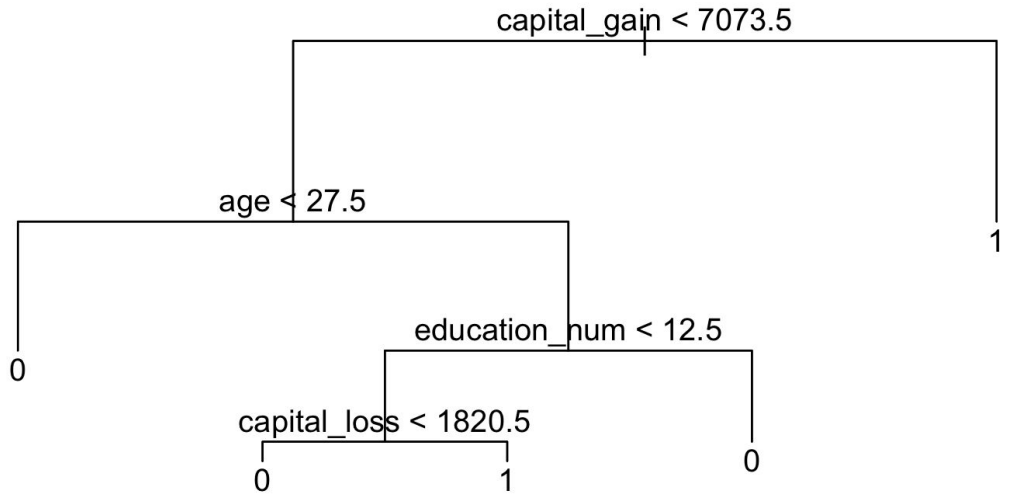
K-means

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 22815  7015
         1  1905   826

               Accuracy : 0.7261
                 95% CI : (0.7212, 0.7309)
    No Information Rate : 0.7592
    P-Value [Acc > NIR] : 1
```

Decision Tree




```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7333 1858
         1   66  512

               Accuracy : 0.8031
                 95% CI : (0.795, 0.8109)
    No Information Rate : 0.7574
    P-Value [Acc > NIR] : < 2.2e-16
```

- I wanted to fit a linear regression model to find the relationship between Annual income and Capital gain, but we do not have information of annual income.
- If I could, then we would know how much you should put into the stock market based on your annual income.
- Last node: high risk high reward.

XGboost

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7021  902
         1  378 1468

               Accuracy : 0.869
                 95% CI : (0.8621, 0.8756)
    No Information Rate : 0.7574
    P-Value [Acc > NIR] : < 2.2e-16
```

Summary

model	accuracy
Baseline	0.7813
Logistic regression with mask	0.7977
K-means	0.7261
KNN	0.7226
Random forest	0.7574
Decision tree	0.8031
xgBoost	0.8690

Thank you

Questions?