Consumer Income Prediction

The Influence of Marital Status and Age on Income

Kailin Kleintjes

About the Dataset

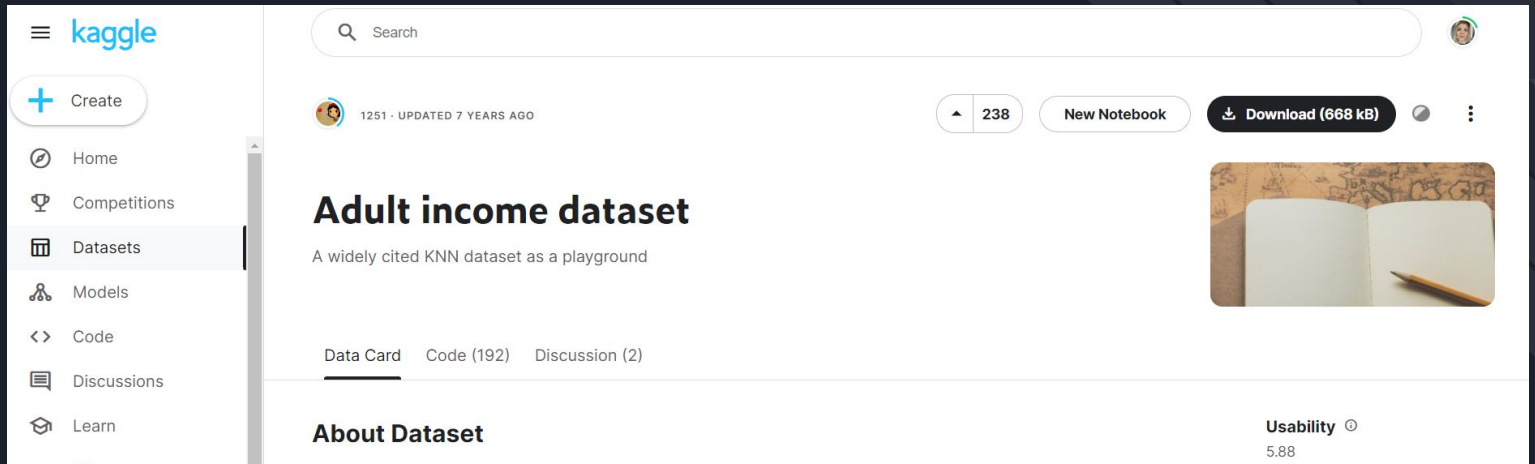
- An individual’s annual income results from various factors. Intuitively, it is influenced by the individual’s education level, age, gender, occupation, and etc.

Objective

- We can explore the possibility in predicting income level based on the individual’s personal information.

Source:

[Adult Income Dataset]



Brief description of data

This data set looks at possible influences on an individual's annual income (e.g. education level, age, gender, occupation, and etc.). The target for this data set is income, specifically if that income is over or under $50,000. This is a classification problem as there are only two answers for our target. Each row represents a person that is evaluated over 14 features.

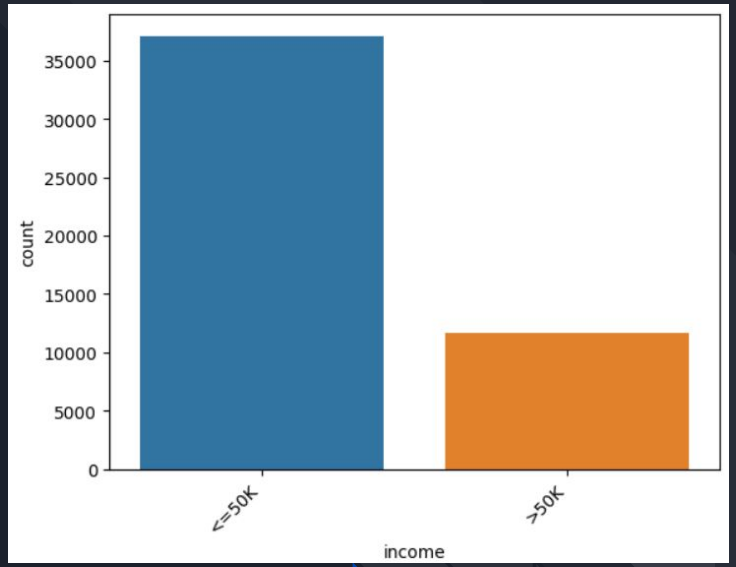
The Stakeholder

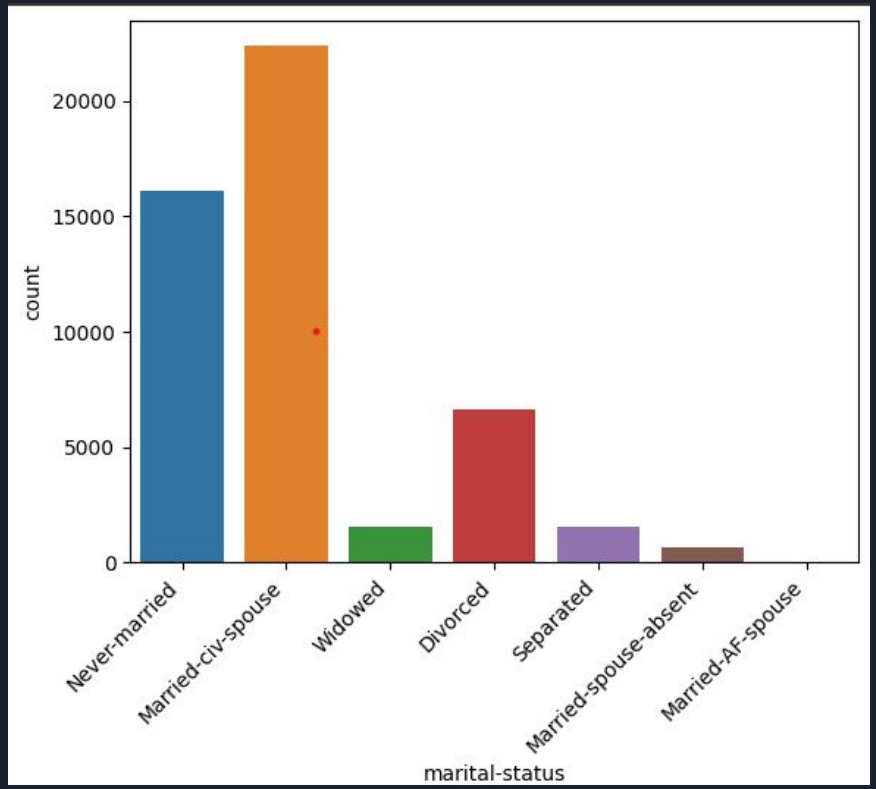
As a business stakeholder this information could be beneficial in determining what population groups to target and with what type of item.

For example if income was higher for one population generally (e.g. older in age) you may offer a luxury version of the item. On the flip side you could use this info to market an economy version to the lower income (e.g. younger) group.

Income

The majority of the sample has an income less than or equal to 50k.





Marital Status

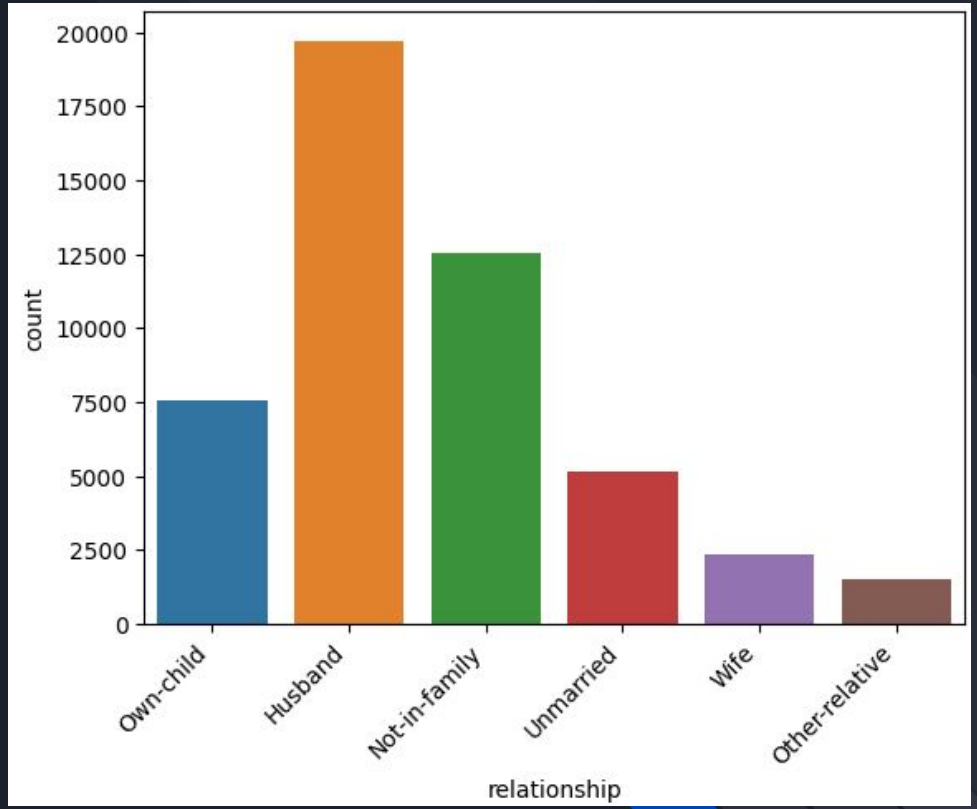
Number of participants in each marital status.

Observation: The most common status is “married civilian (civ) spouse”.

Relationship

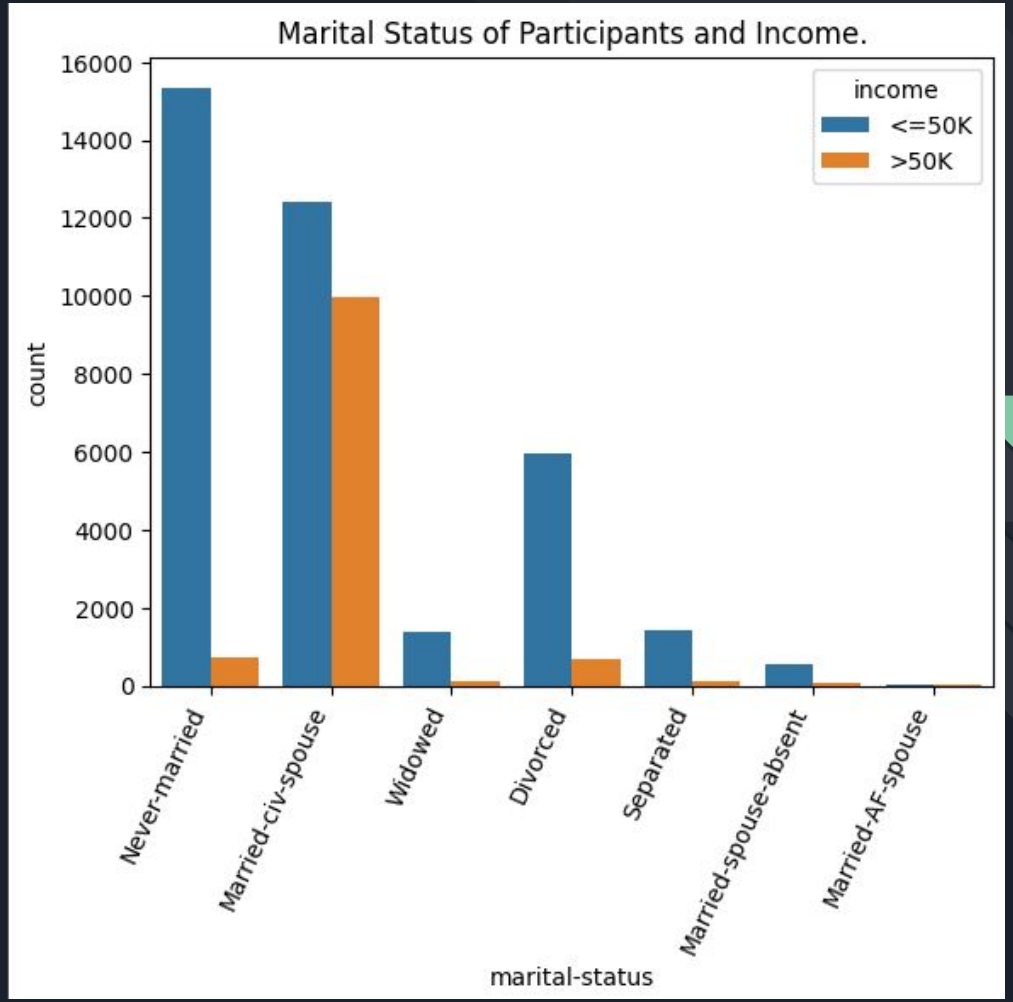
Number of participants in each relationship status.

Observation: The most common relationship is husband.



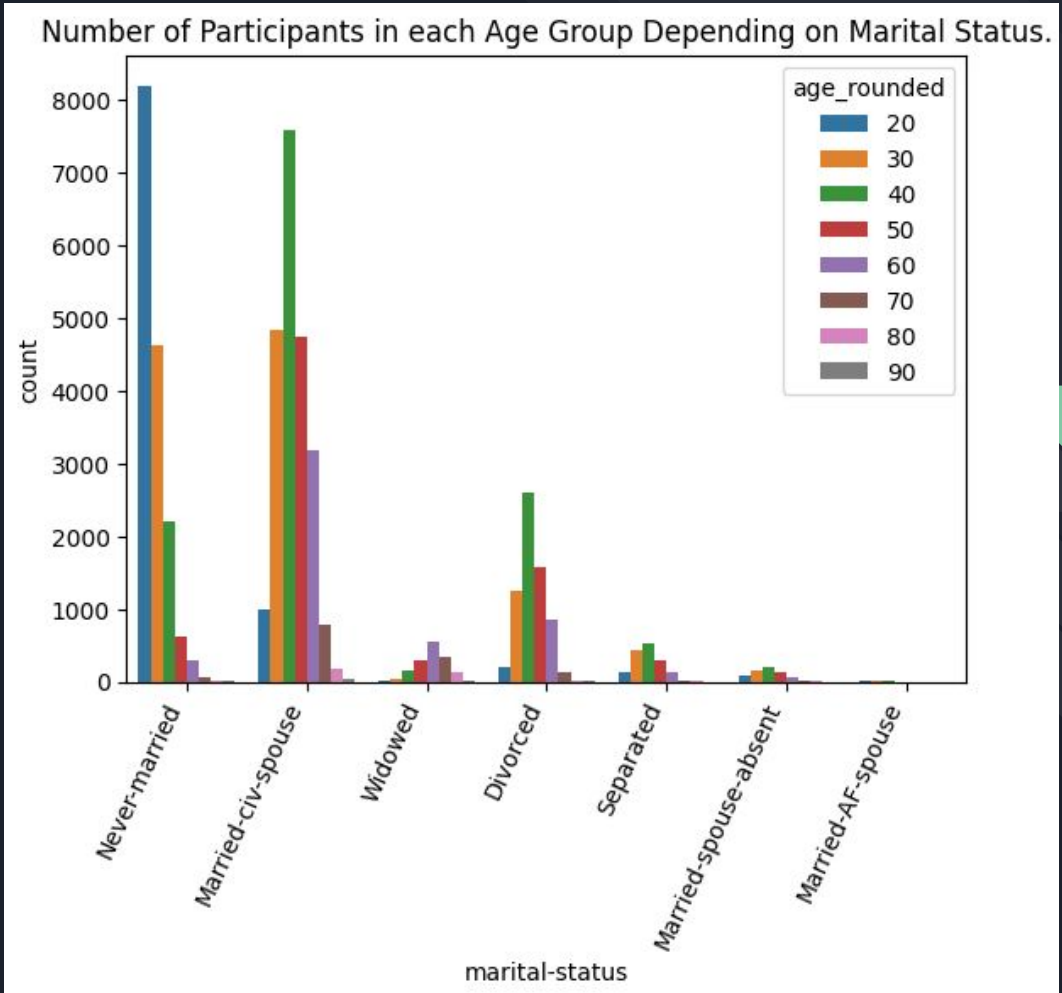
#1

The next visualization shows which income is obtained based on marital status. We can see that a majority of the participants fall in the under / equal to 50k income range. We can also see a significant difference between the category “married civilian spouse" and all other marital categories as it has the highest number of participants making over 50k.


Marital Status of Participants and Income.
income
<=50K
>50K
count
16000
14000
12000
10000
8000
6000
4000
2000
0
Never-married
Married-civ-spouse
Widowed
Divorced
Separated
Married-spouse-absent
Married-AF-spouse
marital-status

#2

These marital groups are further explored in a second visualization in which each type of marital status is separated further into age groups. All categories except for “never married” seem to fall on a bell curve with “never married" skewed to the right.


Number of Participants in each Age Group Depending on Marital Status.
age_rounded
20
30
40
50
60
70
80
90
count
8000
7000
6000
5000
4000
3000
2000
1000
0
Never-married
Married-civ-spouse
Widowed
Divorced
Separated
Married-spouse-absent
Married-AF-spouse
marital-status

What does this tell us about predicting income? Based on these two visualizations if we know a participant's marital status and age we can better predict which income class they fall in. For example a 20 year old who has never been married is most likely to fall in the under/equal to 50K range. However, if we have a 40-year-old who married a civilian spouse we can predict that they are most likely in the over 50k range.

The Model

The model that has been developed shows that we can identify individuals making under <=50k with 87% precision. However we are less capable at identifying individuals making over >50K.

Recommendation

Using this model I would recommend focusing on the lower income crowd as the ability to correctly identify a fitting customer is higher with continued model modification.

Model improvement

This model can be reconfigured in multiple ways that could provide for different results, such as identifying and assigning some feature to ordinal encoding instead of nominal. Additionally taking the balance of the sample into account and dpoint a SMOTE analysis may bring interesting results.