1. How big is the dataset? Write the total number of instances in the dataset.

According to Documentation - 48842

But if do pd.shape this is the dataset size

```
32561 -> Rows 13 -> Coloumns
```

2. What type of dataset? Describe the type of dataset provided (e.g., graph, time series, database records, etc.)

Continuous, Categorical, and Multivariate data containing details of the individuals mapped to several attributes.

3. What is dimensionality? Write the dimensionality of the data instances.

32561 x 13

4. What are the data types of the features? For each feature/attribute, write the data type (e.g., float, string, integer, boolean).

```
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32561 non-null  int64
 1   workclass       30725 non-null  object
 2   education       32561 non-null  object
 3   marital-status  32561 non-null  object
 4   occupation      30718 non-null  object
 5   relationship    32561 non-null  object
 6   race            32561 non-null  object
 7   sex             32561 non-null  object
 8   capital-gain    32561 non-null  int64
 9   capital-loss    32561 non-null  int64
 10  hours-per-week  32561 non-null  int64
 11  native-country  31978 non-null  object
 12  class           32561 non-null  object
```

5. How was the dataset collected?
Describe who collected the dataset, how it was collected, where it was collected from, and when It was collected. Additionally, describe why you think the dataset might have been collected.

The dataset was:

1. **Collected / Extracted by: Barry Becker**
2. **Donated By: Ronny Kohavi and Barry Becker**
3. **How it was collected: A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))**
4. **Collected on: 1994**
5. **Donated on: 1996-05-01**
6. **I think this database was collected to predict and determine whether a person makes over 50K a year.**

6. Are there any sensitive attributes? List any features you think might be sensitive based on privacy or ethics-related issues.

I think sex, native-country, and race are sensitive information.

7. What is the dataset quality? Describe whether you think the entries in the dataset are trustworthy and if there are any quality issues you think might need to be considered.

Data is quite old and stale and might not be applicable to the current time in 2023.
But, I think data is authentic and can be used to understand the general pattern which is applicable at this time as well.

8. What do I want to find?
Describe the goal of the analysis including a) what you are trying to predict, b) how you would measure if your classifier is doing a good job or not for the task, and c) whether there are any limitations you foresee in using the chosen dataset to address Phoenix Solar's problem.

a) What you are trying to predict:
We are trying to find the potential income of people from Dataset, we can use binary classification since there are only two categories, more than $50K and less than $50K.

b) How you would measure if your classifier is doing a good job or not for the task:
We can use metrics such as accuracy, precision, recall, and F1 score to evaluate the performance of the classifier.

1. **Accuracy measures the proportion of correct predictions made by the classifier.**
2. **Precision measures the proportion of true positives among all predicted positives.**
3. **Recall measures the proportion of true positives among all actual positives.**
4. **The F1 score is the harmonic mean of precision and recall.**

We need to choose a suitable metric that best suits our problem.

c) Whether there are any limitations you foresee in using the chosen dataset to address Phoenix Solar's problem:
The chosen dataset may have some limitations that may affect the analysis. For example, the dataset may have missing values, outliers, or noise. These issues may need to be addressed to ensure that the analysis is reliable.
Additionally, the dataset may not contain all the relevant variables needed to predict customer behavior. We need to ensure that the dataset is suitable for our purpose, and if not, we may need to collect more data or find a new dataset.
Also, data is stale and might not be useful in 2023.

9. Are any values missing? Write the number of instances with missing values (if any).

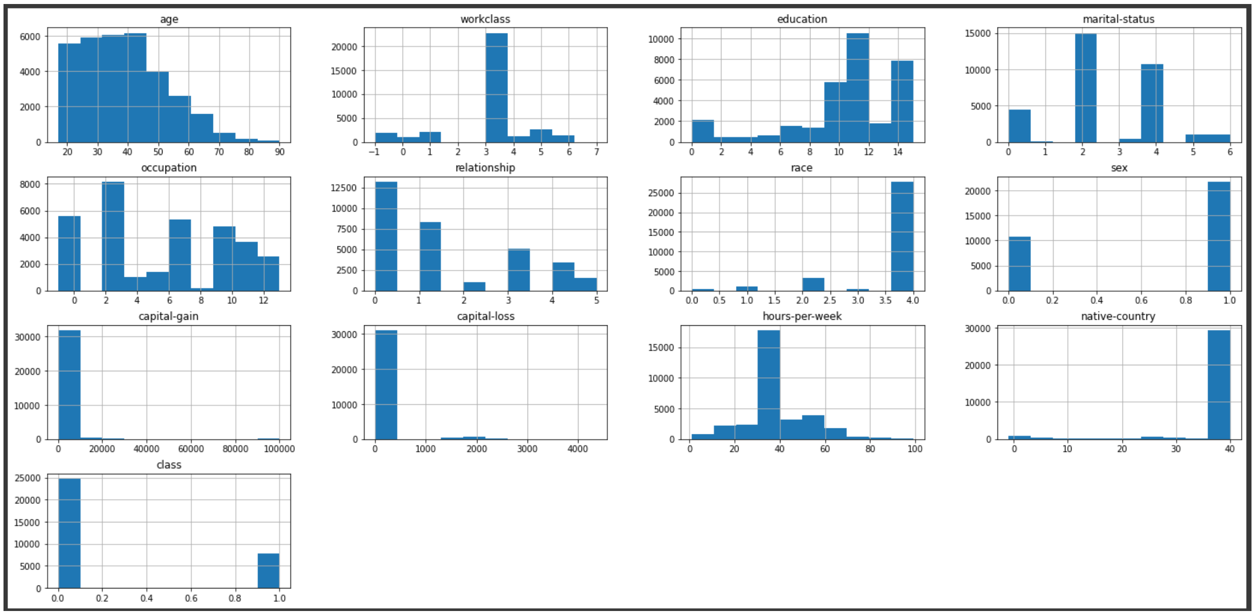
Total Number of Missing Values: 2399

10. Are there any duplicate instances? Write the number of duplicate instances (if any).

Total Number of Duplicate Values: 3465

11. What is the distribution of my attributes? Plot the distribution of each feature you are considering using in the dataset.

Here is the distribution of the attributes.

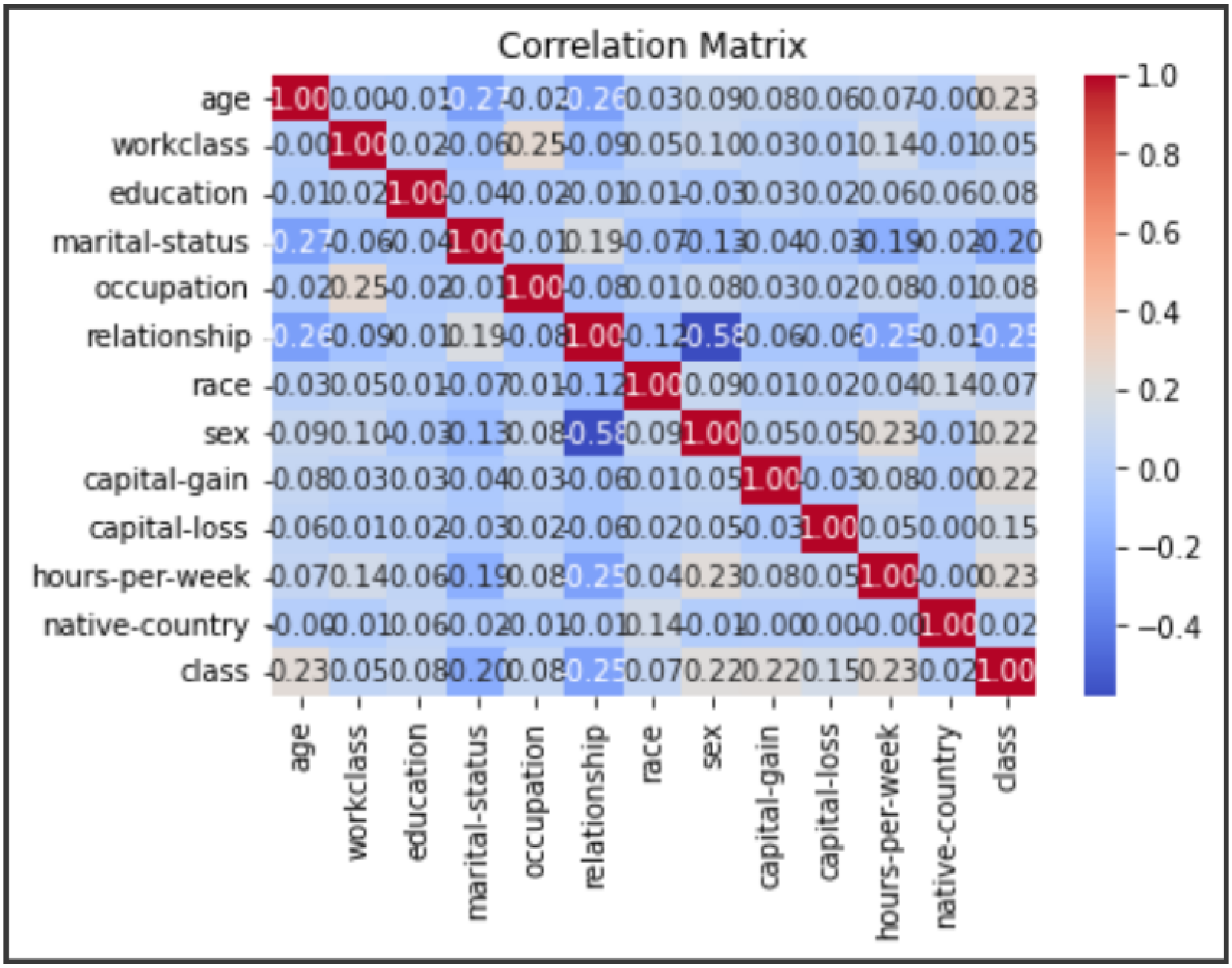


12. What is the distribution of my labels?
Write the number of high-income and low-income instances in the dataset. Note whether there is any class imbalance.

<=50K 24720
>50K 7841
There is a class imbalance, <=50K class has almost 3 times the entries than other one.

13. How are the features related to each other? Compute the correlation coefficient between each pair of features you are considering using in the dataset. State which features are strongly correlated with each other (if any).

I kept the threshold of abs(correlations) > 0.7, and I didn't find any strong correlation among the features. But, I think sex and relationships are pretty close with a score of -0.58.



14. What are the most important features?
Compute the correlation coefficient between each feature you are considering and the class.
State which features are strongly correlated with the target class (if any).

```
class             1.000000
age               0.234037
hours-per-week    0.229689
capital-gain      0.223329
capital-loss      0.150526
education         0.079317
occupation        0.075468
workclass         0.051604
marital-status   -0.199307
relationship     -0.250918
Name: class, dtype: float64
Features strongly correlated with the target class: None except class    1.0
```

I am not considering sensitive features like sex, race, and native-country.

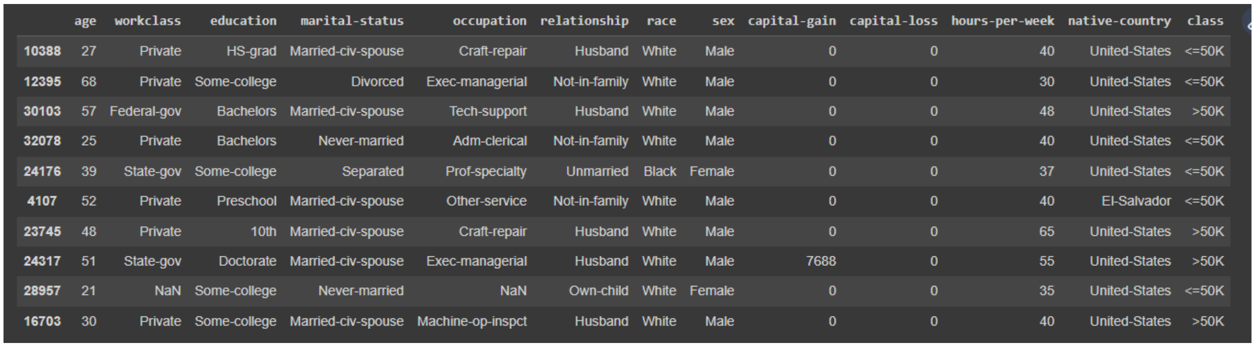
I think, the most important features are: age, hours per week, capital-gain, and relationship.

15. What are the least important features?
State which features have very weak or no correlation with the target class (if any).

Looking at the above scores, education, occupation, workclass are pretty weak compared to class.

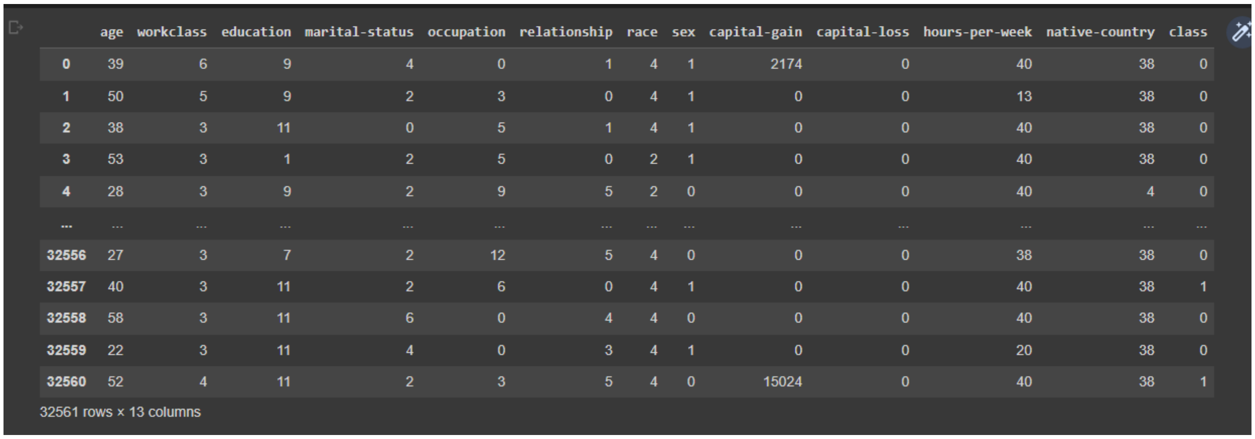
16. What do the samples look like? Visualize some example instances in the dataset (e.g., print the rows).

Here are the sample rows, before and after applying conversions.
Before:



	age	workclass	education	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	class
10388	27	Private	HS-grad	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	<=50K
12395	68	Private	Some-college	Divorced	Exec-managerial	Not-in-family	White	Male	0	0	30	United-States	<=50K
30103	57	Federal-gov	Bachelors	Married-civ-spouse	Tech-support	Husband	White	Male	0	0	48	United-States	>50K
32078	25	Private	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
24176	39	State-gov	Some-college	Separated	Prof-specialty	Unmarried	Black	Female	0	0	37	United-States	<=50K
4107	52	Private	Preschool	Married-civ-spouse	Other-service	Not-in-family	White	Male	0	0	40	El-Salvador	<=50K
23745	48	Private	10th	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	65	United-States	>50K
24317	51	State-gov	Doctorate	Married-civ-spouse	Exec-managerial	Husband	White	Male	7688	0	55	United-States	>50K
28957	21	NaN	Some-college	Never-married	NaN	Own-child	White	Female	0	0	35	United-States	<=50K
16703	30	Private	Some-college	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K

After:



	age	workclass	education	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	class
0	39	6	9	4	0	1	4	1	2174	0	40	38	0
1	50	5	9	2	3	0	4	1	0	0	13	38	0
2	38	3	11	0	5	1	4	1	0	0	40	38	0
3	53	3	1	2	5	0	2	1	0	0	40	38	0
4	28	3	9	2	9	5	2	0	0	0	40	4	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
32556	27	3	7	2	12	5	4	0	0	0	38	38	0
32557	40	3	11	2	6	0	4	1	0	0	40	38	1
32558	58	3	11	6	0	4	4	0	0	0	40	38	0
32559	22	3	11	4	0	3	4	1	0	0	20	38	0
32560	52	4	11	2	3	5	4	0	15024	0	40	38	1

32561 rows × 13 columns

17. Do I have all the information I need?
Describe whether there are any other relevant variables that are not captured in the chosen dataset. These could be attributes that you suggest to Phoenix Solar they should try to collect, although you do not have them available for this assignment.

I think we have most of the attributes that are necessary, but I think, the company name, revenue, and capabilities of the person would have had a better impact on model predictions.

18. For whom/what purpose am I mining this dataset?
Describe the purpose of your data mining analysis and how your solution will meet (or not meet) the needs of Phoenix solar.

I am mining this dataset for a company called Phoenix Solar, as they hired me to help them estimate how many people in Phoenix have low vs high income.

My solution has 2 models trained on the dataset provided. Before training the model I cleaned the dataset and dropped the sensitive information that is not useful for training the model and acquired an accuracy of 83.82% using the Decision Tree Classifier.

I have used SVM as well, to compare the results.