**Assignment 1**

**Titanic- Machine Learning from Disaster**

**Hongyi Hao**



**1 Problem Description**

* 1. Problem 1:

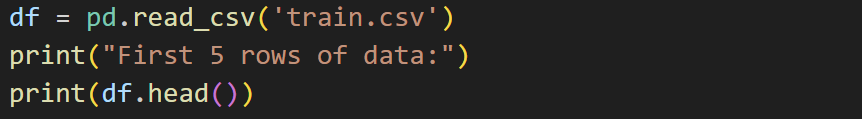
Which feature do you believe is the most important for the model’s performance? Justify your answer with evidence, such as data analysis, visualizations, or feature importance scores.

* 1. Problem 2:

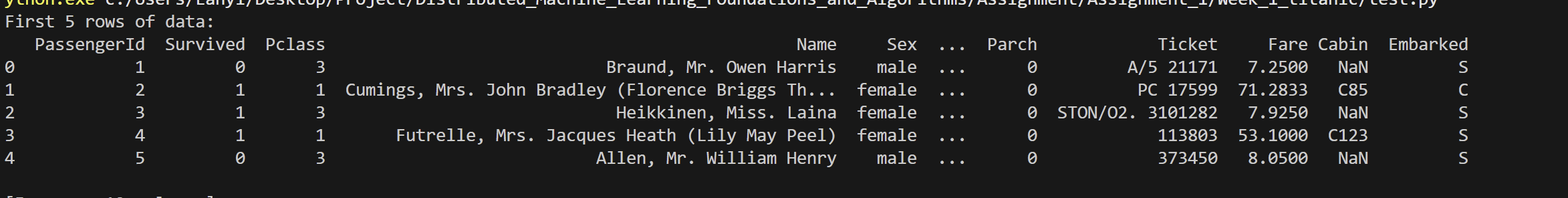
For the "Age" feature, what alternative methods could be used to handle missing values instead of filling them with 0? Explain how these methods could improve the model’s performance.

1. **Data Preparation**
   1. Read and Visualize Data
2. Read the train.csv file.
3. Print out the first 5 samples.

To begin with, I used the ***pandas.read\_csv*** function in Python to read the train.csv data. And then print out the first 5 samples.



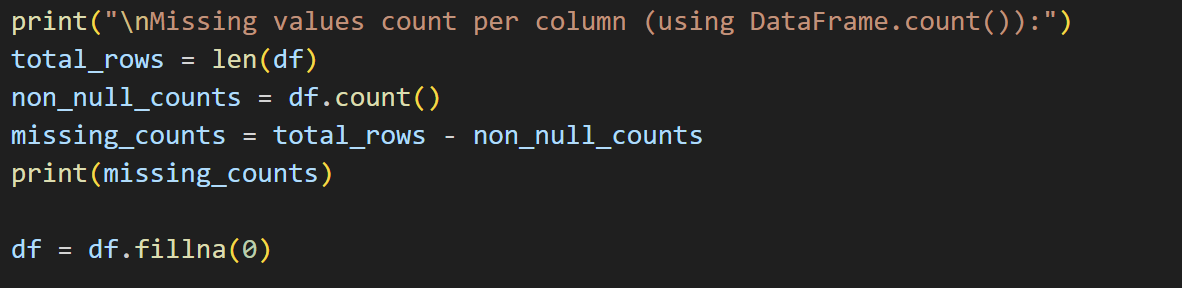
**Output:**



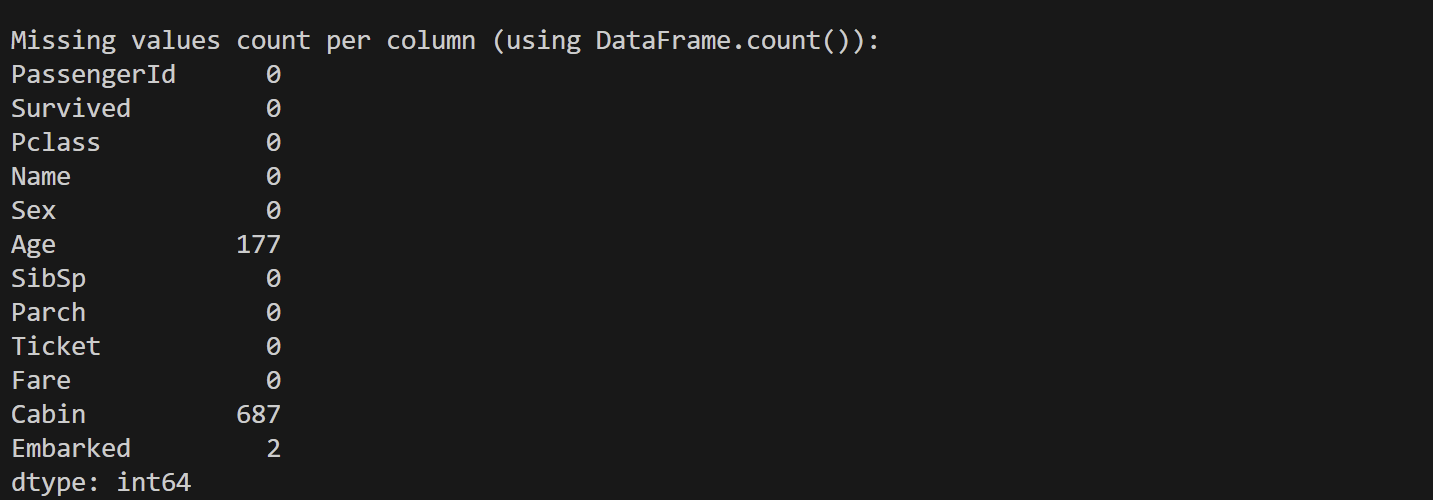
* 1. Handling Missing Data

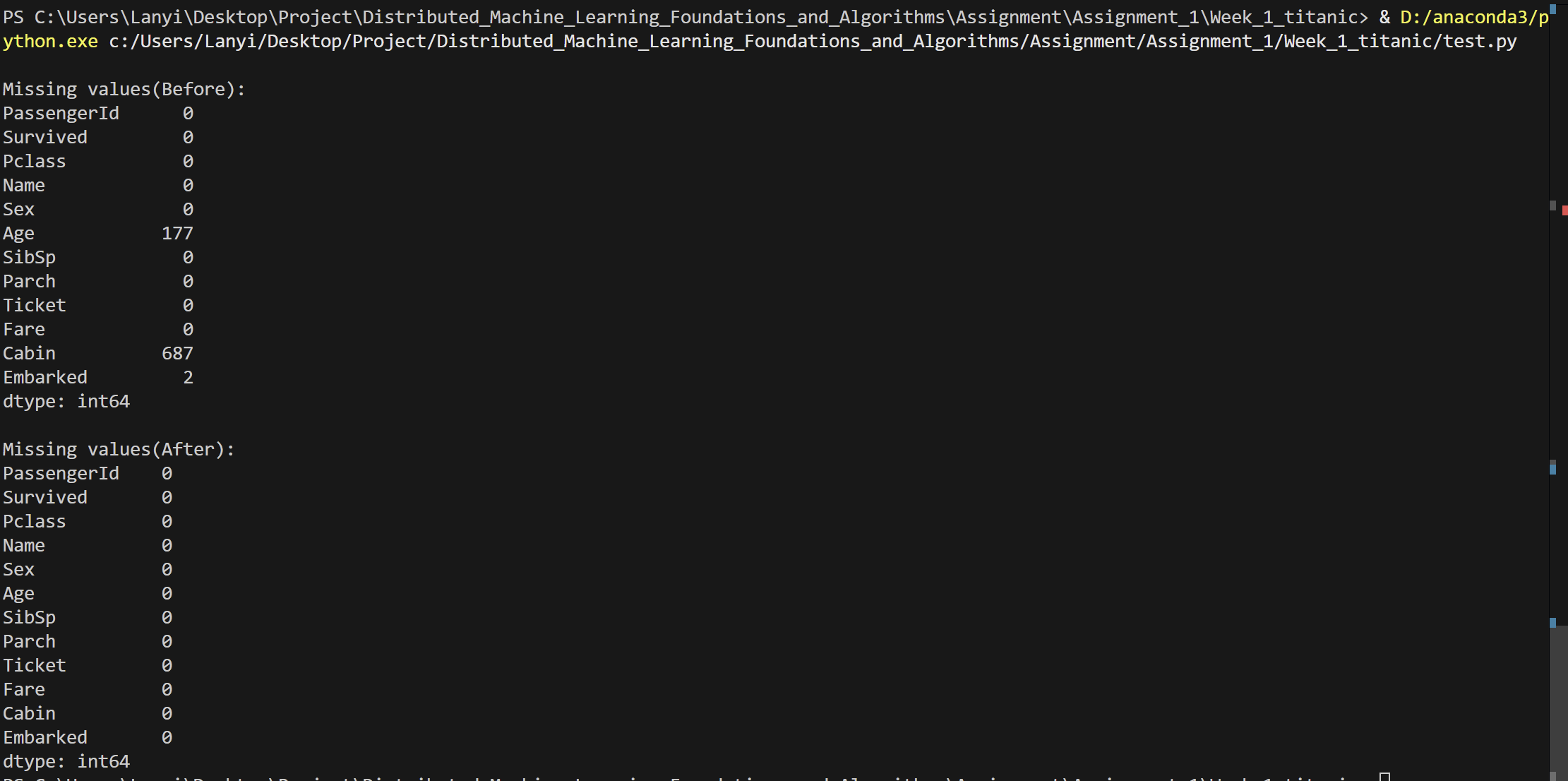
1. Print out the count of missing values in each column.
2. Impute missing entries with zero.

Then, I used the ***pandas.DataFrame.count*** and ***pandas.DataFrame.fillna*** function to count the number of values in each column and impute missing entries with 0.



**Output:**

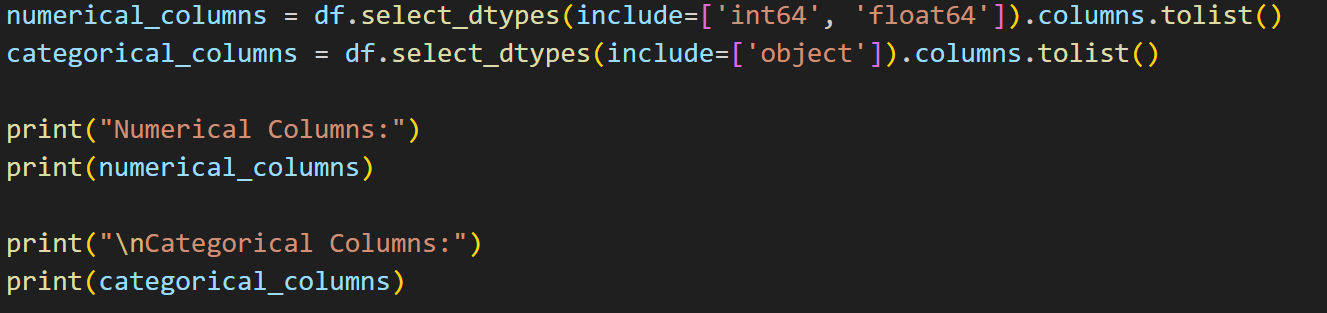


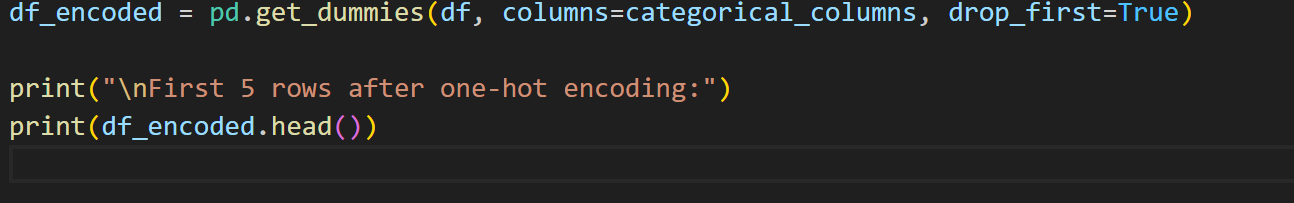
**Comparison Before And After Filling:**

* 1. Handling numerical and categorical values.

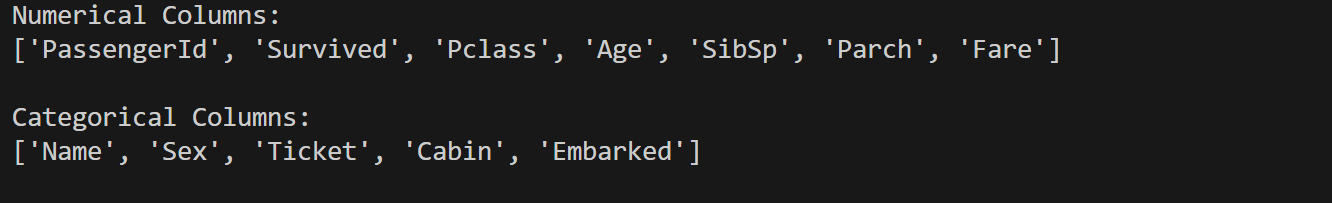
1. Identify and write down which columns are numerical and which are categori cal.
2. Convert the categorical columns into one-hot encoded columns. Print out the first 5 samples to verify.

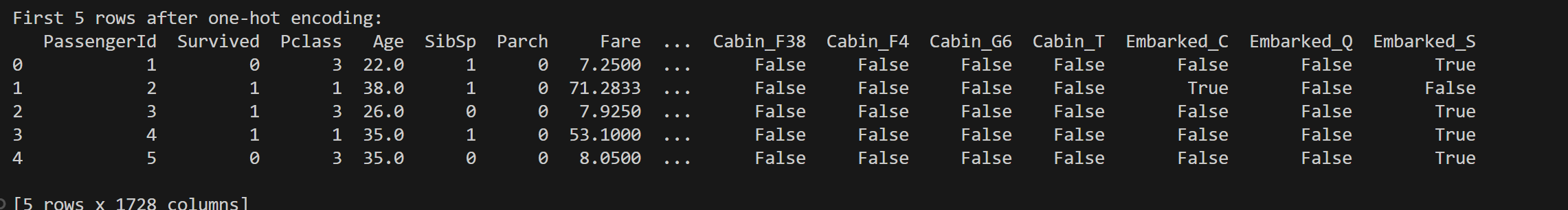
Additionally, I used the ***pandas.DataFrame.head*** function and ***pandas.get\_dummies*** function to Identify and write down the type of the values convert the categorical columns into one-hot encoded columns.





**Output:**





* 1. Dataset Splitting
  2. Addressing Imbalanced Labels for Training Set