2025 USA-NA-AIO Round 1, Problem 3, Part 6

USAAIO **Mar 2025**

Part 6 (5 points, coding and conceptual reasoning task)

In `df_5`, columns `Sex` and `Embarked` are categorical data.

Do the following tasks to process these categorical data.

1. To do logistic regression on this dataset, we need to do one hot encoding on these two columns. Explain why?
2. Do one hot encoding on these two columns. Set `drop_first = True` and `dtype = np.int8`. Save the new dataframe object as `df_6`.
3. Explain what `drop_first = True` means and why we do so.
4. Print the first five rows of `df_5` and `df_6`.
5. Print the shapes of `df_5` and `df_6`.

last visit

USAAIO **Mar 2025**

```
### WRITE YOUR SOLUTION HERE ###

# Question 1
"""
Answer:

Logistic regression requires numerical data, not categorical data.

"""
```

Skip to main content

```
# Question 2
# Answer: (put your code here)

df_6 = pd.get_dummies(df_5, columns=['Sex', 'Embarked'], drop_first = True, dtype

# Question 3
"""
Answer:

Suppose a categorical variable takes value k chosen from K categories, indexed as

By setting drop_first = True, it is replaced by a vector with shape K-1.

If k = 0, then in this vector, all entries are 0.

If k is not 0, then in this vector, the (k-1)th entry (entry indices starts from

Setting drop_first = True avoids multicollinearity.

"""

# Question 4
# Answer: (put your code here)

print(df_5.head())
print(df_6.head())

# Question 5
# Answer: (put your code here)

print(df_5.shape)
print(df_6.shape)

""" END OF THIS PART """
```

Skip to main content

We need numerical values for the input, and one hot encoding avoids assigning an order to the categories.

Related topics

Topic	Replies	Activity
2026 USAAIO Round 1 Sample problems, Problem 9	1	Jan 14
2025 USA-NA-AIO Round 1, Problem 3, Part 2	2	Dec 2025
2025 USA-NA-AIO Round 1, Problem 3, Part 7	1	Mar 2025
2025 USA-NA-AIO Round 1, Problem 3, Part 5	1	Mar 2025
2025 USA-NA-AIO Round 1, Problem 3, Part 1	2	Dec 2025