

02_dataprep_01: Preparing Data for Analysis (Modified Titanic)

The file “titanic_to_student.csv” contains the data with 12 columns.

1. **PassengerId:** A unique identifier assigned to each passenger.
2. **Survived:** Indicates whether the passenger survived (1) or did not survive (0).
3. **Pclass:** The passenger's class (1 = 1st class, 2 = 2nd class, 3 = 3rd class).
4. **Name:** The full name of the passenger.
5. **Sex:** The gender of the passenger (male or female).
6. **Age:** The age of the passenger in years.
7. **SibSp:** The number of siblings or spouses the passenger had aboard the Titanic.
8. **Parch:** The number of parents or children the passenger had aboard the Titanic.
9. **Ticket:** The ticket number assigned to the passenger.
10. **Fare:** The fare paid by the passenger.
11. **Cabin:** The cabin number where the passenger stayed (if known).
12. **Embarked:** The port where the passenger embarked (C = Cherbourg; Q = Queenstown; S = Southampton).

Problems

Problem 1:

How many rows are there in the “titanic_to_student.csv”?

Problem 2:

Drop unqualified variables

Drop variables with missing > 50%

Drop categorical variables (**i.e., all columns except 'Age' and 'Fare'**) with flat values > 70%
(variables with the same value in the same column)

How many columns do we have left?

Note:

-Ensure missing values are considered in your calculation. If you use `normalize` in `.value_counts()`, please include `dropna=False`.

-‘Parch’ and ‘SibSp’ should be considered as numerical variables. (However, for the grader submission, considering only ‘Age’ and ‘Fare’ as numerical variables is fine.)

Problem 3:

Remove all rows with missing targets (the variable "Survived")

How many rows do we have left?

Problem 4:

Handle outliers

For the variable “Fare”, replace outlier values with the boundary values

If value < (Q1 - 1.5IQR), replace with (Q1 - 1.5IQR)

If value > (Q3 + 1.5IQR), replace with (Q3 + 1.5IQR)

What is the average (mean) of “Fare” after replacing the outliers (round 2 decimal points)?

Hint: Use function round(_, 2)

Problem 5:

Impute missing value

For number type column, impute missing values with mean

What is the average (mean) of “Age” after imputing the missing values (round 2 decimal points)?

Hint: Use function round(_, 2)

Problem 6:

Convert categorical to numeric values

For the variable “Embarked”, perform the dummy coding.

What is the average (mean) of “Embarked_Q” after performing dummy coding (round 2 decimal points)?

Hint: Use function round(_, 2)

Problem 7:

Partition data

Split train/test split with stratification using 70%:30% and random seed with 123

Show a proportion between survived (1) and died (0) in all data sets (total data, train, test)

What is the proportion of survivors (survived = 1) in the training data (round 2 decimal points)?

Hint: Use function round(_, 2), and train_test_split() from sklearn.model_selection

Expected Results

Input	Output
Q1	445
Q2	10
Q3	432
Q4	26.27
Q5	29.14
Q6	0.06
Q7	0.41

** Disclaimer: The data used in the example, 'titanic_to_student.csv', differs from the data used for scoring.

Template codes

The template code snippet here is the same as the 'student.py' file included in the attachment to the student file downloaded for this assignment.

```
import pandas as pd
from sklearn.model_selection import train_test_split

"""
    ASSIGNMENT 2 (STUDENT VERSION):
    Using pandas to explore Titanic data from Kaggle (titanic_to_student.csv) and
    answer the questions. (Note that the following functions already take the Titanic
    dataset as a DataFrame, so you don't need to use read_csv.)
"""

def Q1(df):
    """
        Problem 1:
        How many rows are there in the "titanic_to_student.csv"?
    """
    # TODO: Code here
    return None

def Q2(df):
    """
        Problem 2:
        Drop unqualified variables
        Drop variables with missing > 50%
        Drop categorical variables with flat values > 70% (variables with the same
        value in the same column)
        How many columns do we have left?
    """
    # TODO: Code here
    return None

def Q3(df):
    """
        Problem 3:
        Remove all rows with missing targets (the variable "Survived")
    """
```

```

    ...    How many rows do we have left?
    ...
    # TODO: Code here
    return None

def Q4(df):
    ...
    Problem 4:
        Handle outliers
        For the variable "Fare", replace outlier values with the boundary values
        If value < (Q1 - 1.5IQR), replace with (Q1 - 1.5IQR)
        If value > (Q3 + 1.5IQR), replace with (Q3 + 1.5IQR)
        What is the mean of "Fare" after replacing the outliers (round 2 decimal
points)?
        Hint: Use function round(_, 2)
    ...
    # TODO: Code here
    return None

def Q5(df):
    ...
    Problem 5:
        Impute missing value
        For number type column, impute missing values with mean
        What is the average (mean) of "Age" after imputing the missing values
(round 2 decimal points)?
        Hint: Use function round(_, 2)
    ...
    # TODO: Code here
    return None

def Q6(df):
    ...
    Problem 6:
        Convert categorical to numeric values
        For the variable "Embarked", perform the dummy coding.
        What is the average (mean) of "Embarked_Q" after performing dummy coding
(round 2 decimal points)?
        Hint: Use function round(_, 2)
    ...
    # TODO: Code here
    return None

def Q7(df):
    ...
    Problem 7:
        Drop row that contains missing values of "Survived"
        Split train/test split with stratification using 70%:30% and random seed
with 123
        Show a proportion between survived (1) and died (0) in all data sets (total
data, train, test)
        What is the proportion of survivors (survived = 1) in the training data
(round 2 decimal points)?
        Hint: Use function round(_, 2), and train_test_split() from
sklearn.model_selection
    ...
    # TODO: Code here
    return None

```

