04_Titanic_Dataset_Preparation

September 4, 2021

```
[1]: import pandas as pd
     import numpy as np
```

```
[3]: data = pd.read_csv('./TitanicRawDataset.csv')
     data.head()
```

```
[3]:    pclass  survived                                             name     sex  \
     0       1         1                    Allen, Miss. Elisabeth Walton  female
     1       1         1                   Allison, Master. Hudson Trevor    male
     2       1         0                     Allison, Miss. Helen Loraine  female
     3       1         0             Allison, Mr. Hudson Joshua Creighton    male
     4       1         0  Allison, Mrs. Hudson J C (Bessie Waldo Daniels)  female

           age  sibsp  parch  ticket      fare    cabin embarked boat body  \
     0      29      0      0   24160  211.3375       B5        S    2    ?
     1  0.9167      1      2  113781    151.55  C22 C26        S   11    ?
     2       2      1      2  113781    151.55  C22 C26        S    ?    ?
     3      30      1      2  113781    151.55  C22 C26        S    ?  135
     4      25      1      2  113781    151.55  C22 C26        S    ?    ?

                              home.dest
     0                     St Louis, MO
     1  Montreal, PQ / Chesterville, ON
     2  Montreal, PQ / Chesterville, ON
     3  Montreal, PQ / Chesterville, ON
     4  Montreal, PQ / Chesterville, ON
```

```
[4]: data = data.replace('?', np.nan)
     data.isnull().sum()
```

```
[4]: pclass        0
     survived      0
     name          0
     sex           0
     age         263
     sibsp         0
     parch         0
```

```
ticket          0
fare            1
cabin        1014
embarked        2
boat          823
body         1188
home.dest     564
dtype: int64
```

[5]:

```python
def get_first_cabin(row):
    try:
        return row.split()[0]
    except:
        return np.nan
```

[6]:

```python
data['cabin'] = data['cabin'].apply(get_first_cabin)
```

[7]:

```python
data.to_csv('./titanic.csv', index=False)
```