Identifying numerical and categorical variables

Numerical variables can be discrete or continuous.

Numerical Variables

Discrete variables are those where the pool of possible values is finite and are generally whole numbers such as 1, 2, and 3.

Examples for Discreate Variables: number of children, number of pets, or the number of bank accounts.

Continuous variables are those whose values may take any number within a range.

Examples for Continuous Variables: the price of a product,income, house price, or interest rate.

Categorical variables

Categorical Variables are values that are selected from a group of categories, also called labels.

Examples of Categorical Variables: gender, which takes values of male and female, or country of birth, which takes values of Argentina, Germany, and so on.

```
# Load the libraries that are required
import pandas as pd
import matplotlib.pyplot as plt
```

```
# Load the titanic dataset and inspect the variables.
data = pd.read_csv('data/titanic.csv')
data.dtypes
```

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

Note: Discrete variables are usually of the **int** type, continuous variables are usually of the **float** type, and categorical variables are usually of the **object** type when they're stored in pandas.

observations: Here we can observe PassengerId,Survived,Pclass,SibSp,Parch are **Int type** then Name,Sex,Ticket,Cabin,Embarked are of **object type** and Age,Fare are of **Float type**.

Tip: In many datasets, integer variables are cast as float . So, after inspecting the data type of the variable, even if you get float as output, go ahead and check the unique values to make sure that those variables are discrete and not continuous.

```
# Inspect the distinct values of the sibsp discrete variable
data['SibSp'].unique()
```

```
array([1, 0, 3, 4, 2, 5, 8])
```

```
# Now, let's inspect the first 20 distinct values of the continuous
variable fare
data['Fare'].unique()[:20]
```

```
array([ 7.25  , 71.2833,  7.925 , 53.1   ,  8.05  ,  8.4583, 51.8625,
       21.075 , 11.1333, 30.0708, 16.7   , 26.55  , 31.275 ,  7.8542,
       16.    , 29.125 , 13.    , 18.    ,  7.225 , 26.    ])
```

```
# inspect the Unique values of Embarked feature - Categorical Variable
data['Embarked'].unique()
```

```
array(['S', 'C', 'Q', nan], dtype=object)
```

```
# inspect the Unique values of Cabin feature - Mixed Variable
data['Cabin'].unique()
```

```
array([nan, 'C85', 'C123', 'E46', 'G6', 'C103', 'D56', 'A6',
       'C23 C25 C27', 'B78', 'D33', 'B30', 'C52', 'B28', 'C83', 'F33',
       'F G73', 'E31', 'A5', 'D10 D12', 'D26', 'C110', 'B58 B60',
'E101',
       'F E69', 'D47', 'B86', 'F2', 'C2', 'E33', 'B19', 'A7', 'C49',
'F4',
       'A32', 'B4', 'B80', 'A31', 'D36', 'D15', 'C93', 'C78', 'D35',
       'C87', 'B77', 'E67', 'B94', 'C125', 'C99', 'C118', 'D7', 'A19',
       'B49', 'D', 'C22 C26', 'C106', 'C65', 'E36', 'C54',
       'B57 B59 B63 B66', 'C7', 'E34', 'C32', 'B18', 'C124', 'C91',
'E40',
       'T', 'C128', 'D37', 'B35', 'E50', 'C82', 'B96 B98', 'E10',
'E44',
       'A34', 'C104', 'C111', 'C92', 'E38', 'D21', 'E12', 'E63',
'A14',
       'B37', 'C30', 'D20', 'B79', 'E25', 'D46', 'B73', 'C95', 'B38',
       'B39', 'B22', 'C86', 'C70', 'A16', 'C101', 'C68', 'A10', 'E68',
       'B41', 'A20', 'D19', 'D50', 'D9', 'A23', 'B50', 'A26', 'D48',
       'E58', 'C126', 'B71', 'B51 B53 B55', 'D49', 'B5', 'B20', 'F
G63',
       'C62 C64', 'E24', 'C90', 'C45', 'E8', 'B101', 'D45', 'C46',
'D30',
       'E121', 'D11', 'E77', 'F38', 'B3', 'D6', 'B82 B84', 'D17',
```

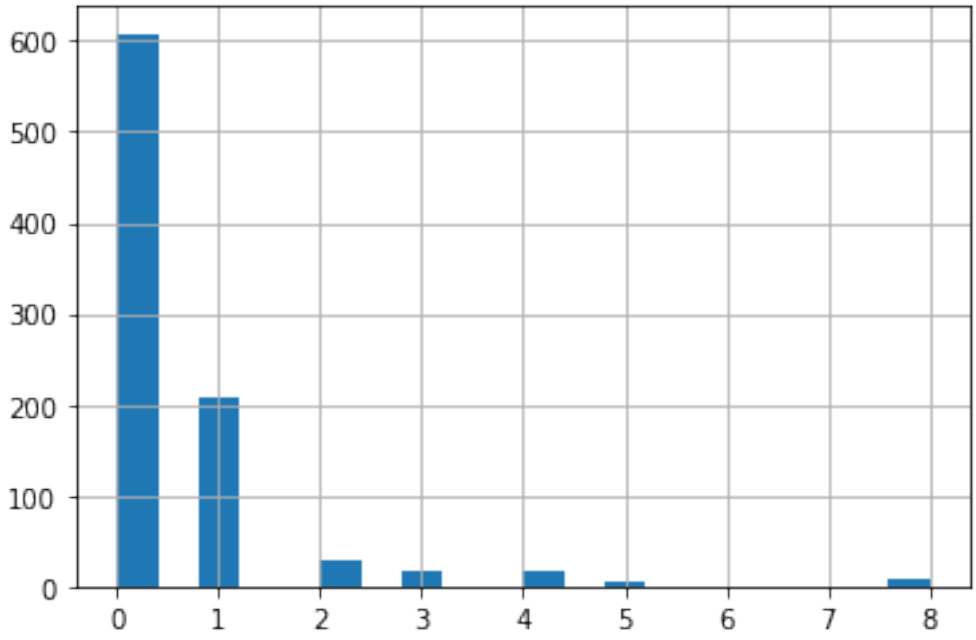
```
'A36',
       'B102', 'B69', 'E49', 'C47', 'D28', 'E17', 'A24', 'C50', 'B42',
       'C148'], dtype=object)
```

Note: The embarked variable contains strings as values, which means it's categorical, whereas cabin contains a mix of letters and numbers, which means it can be classified as a mixed type of variable.

Visualizations

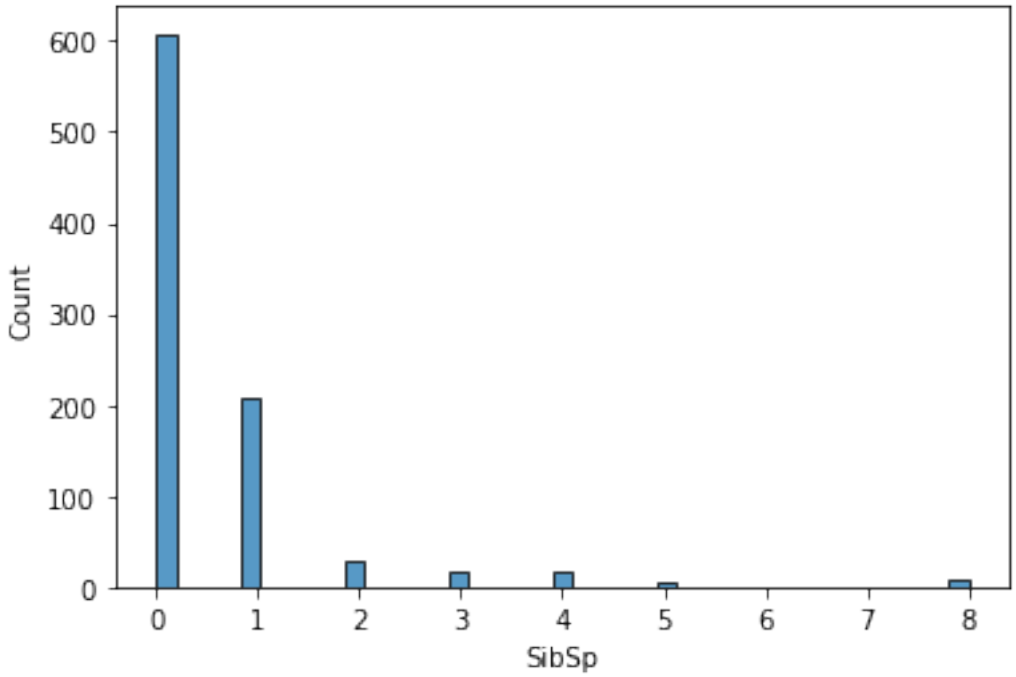
```
# Let's make a histogram for the sibsp variable by dividing the
variable value range into 20 intervals

data['SibSp'].hist(bins=20)
```

```
<AxesSubplot:>
```



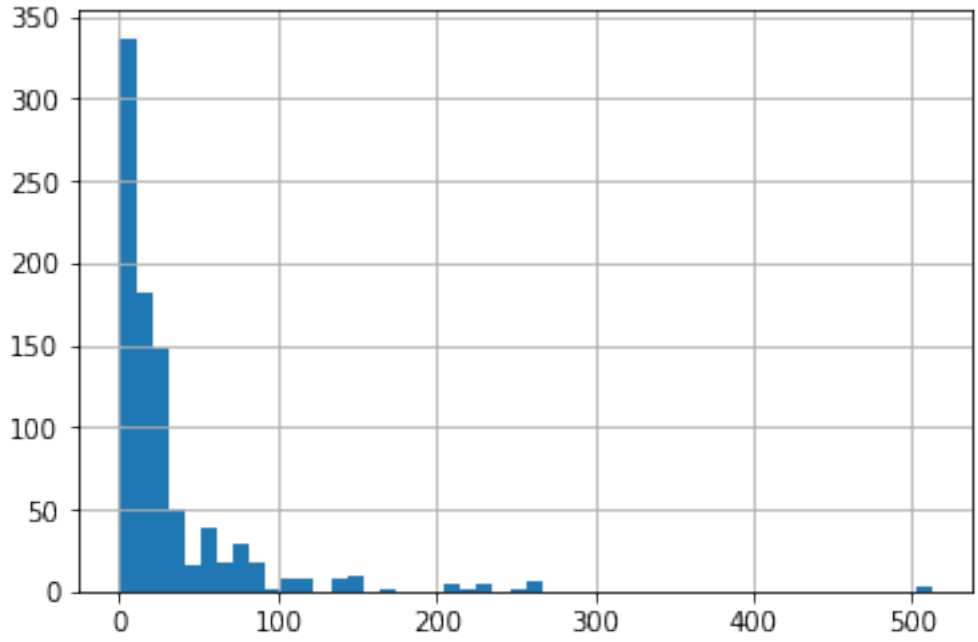

```
# we can also plot this using seaborn
import seaborn as sns
sns.histplot(data['SibSp'])
```

```
<AxesSubplot:xlabel='SibSp', ylabel='Count'>
```



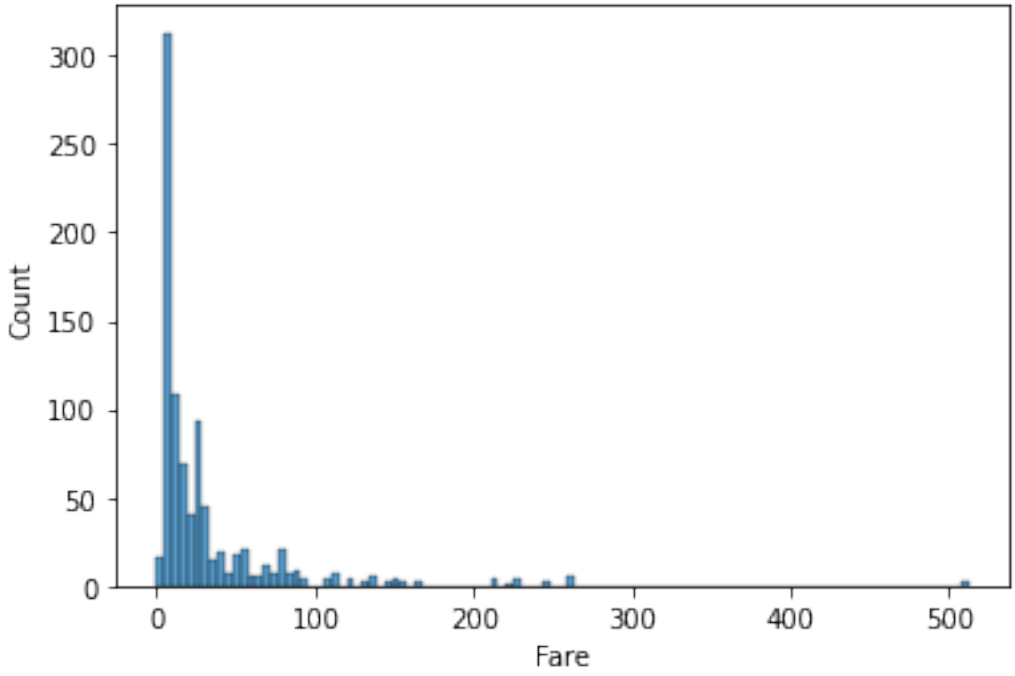

```
# let's make a histogram of the fare variable by sorting the values
into 50 contiguous intervals

data['Fare'].hist(bins=50)
```

```
<AxesSubplot:>
```




```
# by using seaborn
sns.histplot(data['Fare'])
```

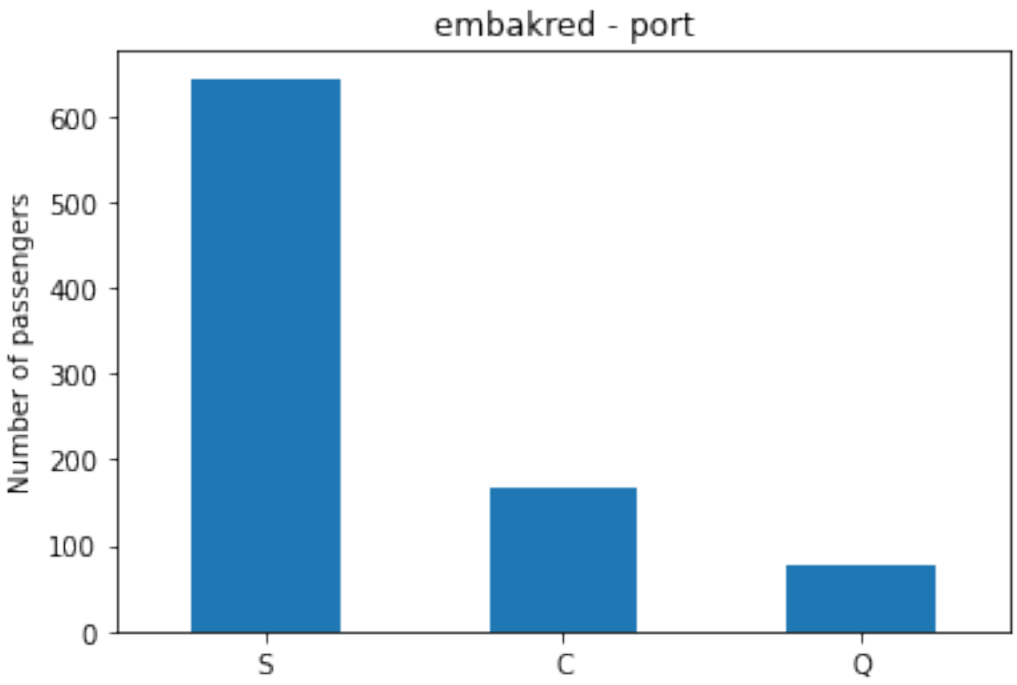
```
<AxesSubplot:xlabel='Fare', ylabel='Count'>
```



Note: The histogram of continuous variables shows values throughout the variable value range.

```
# bar plots for categorical Variable
data['Embarked'].value_counts().plot.bar()
plt.xticks(rotation=0)
plt.ylabel('Number of passengers')
plt.title('embakred - port')
```

```
Text(0.5, 1.0, 'embakred - port')
```


embakred - port
Number of passengers
600
500
400
300
200
100
0
S
C
Q