NAME : ARYAN SIRDESAI ROLL No. : TACO20175 Lab Assignment 1 : Data Wrangling I

Perform the following operations using Python on any open source dataset (e.g., data.csv)

1. Import all the required Python Libraries.
2. Locate an open source data from the web (e.g., https://www.kaggle.com). Provide a clear description of the data and its source (i.e., URL of the web site).
3. Load the Dataset into pandas dataframe.
4. Data Preprocessing: check for missing values in the data using pandas isnull(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.
5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.
6. Turn categorical variables into quantitative variables in Python.

In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [ ]:
```
5+3
```

Out[ ]:
```
8
```

In [ ]:
```
print("Aryan Sirdesai");
```

```
Aryan Sirdesai
```

In [ ]:
```
df = pd.read_csv('train.csv')
```

In [ ]:
```
df
```

Out[ ]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	N
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	N
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	N
...	...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	N
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	N
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	N

891 rows × 12 columns

In [ ]:
```
df.head()
```

Out[ ]:

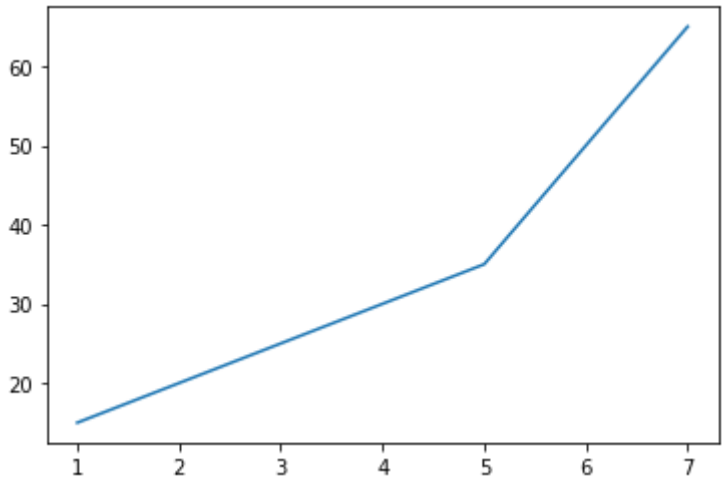
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN

In [ ]:
```
df.describe()
```

Out[ ]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [ ]:
```
x = [1,3,5,7]
y = [15,25,35,65]
plt.plot(x,y)
plt.show()
```



In [ ]:
```
df.isnull()
```

Out[ ]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embar
0	False	False	False	False	False	False	False	False	False	False	True	F
1	False	False	False	False	False	False	False	False	False	False	False	F
2	False	False	False	False	False	False	False	False	False	False	True	F
3	False	False	False	False	False	False	False	False	False	False	False	F
4	False	False	False	False	False	False	False	False	False	False	True	F
...	...	...	...	...	...	...	...	...	...	...	...	
886	False	False	False	False	False	False	False	False	False	False	True	F
887	False	False	False	False	False	False	False	False	False	False	False	F
888	False	False	False	False	False	True	False	False	False	False	True	F
889	False	False	False	False	False	False	False	False	False	False	False	F
890	False	False	False	False	False	False	False	False	False	False	True	F

891 rows × 12 columns

In [ ]:
```
import seaborn as sns
```

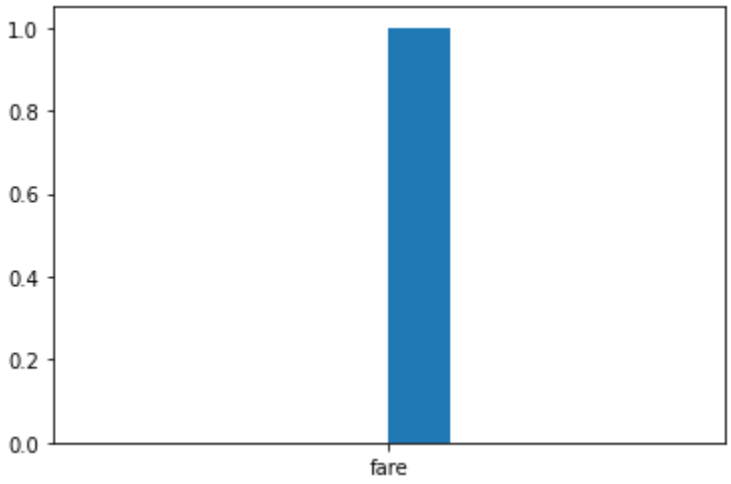
In [ ]:
```
x = [1,3,5,7]
y = [15,25,35,65]
plt.histogram(x,y)
plt.histogram()
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-12-1a02443b34e7> in <module>
      1 x = [1,3,5,7]
      2 y = [15,25,35,65]
----> 3 plt.histogram(x,y)
      4 plt.histogram()

AttributeError: module 'matplotlib.pyplot' has no attribute 'histogram'
```

In [ ]:
```
x=['fare']
```

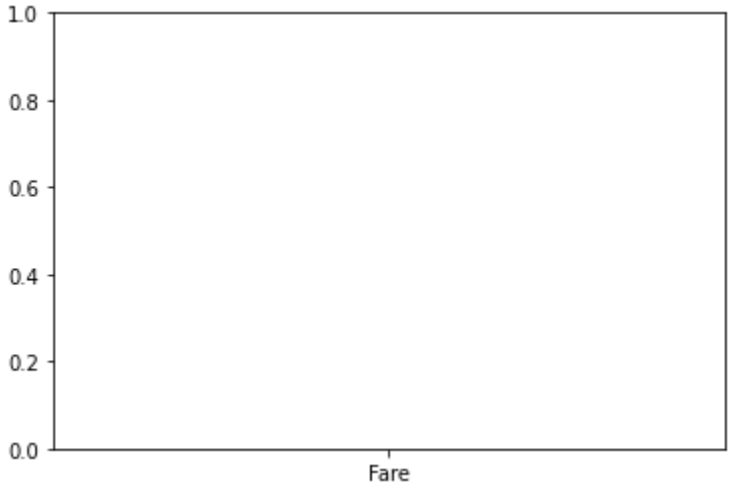
```
plt.hist(x)
```

Out[ ]:
```
(array([0., 0., 0., 0., 0., 1., 0., 0., 0., 0.]),
 array([-0.5, -0.4, -0.3, -0.2, -0.1,  0. ,  0.1,  0.2,  0.3,  0.4,  0.5]),
 <BarContainer object of 10 artists>)
```



In [ ]:
```
plt.hist('Fare','Age')
plt.show()
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-21-3d241efc0eaf> in <module>
----> 1 plt.hist('Fare','Age')
      2 plt.show()
      3

/usr/local/lib/python3.9/dist-packages/matplotlib/pyplot.py in hist(x, bins, rang
e, density, weights, cumulative, bottom, histtype, align, orientation, rwidth, lo
g, color, label, stacked, data, **kwargs)
   2598         orientation='vertical', rwidth=None, log=False, color=None,
   2599         label=None, stacked=False, *, data=None, **kwargs):
-> 2600     return gca().hist(
   2601         x, bins=bins, range=range, density=density, weights=weights,
   2602         cumulative=cumulative, bottom=bottom, histtype=histtype,

/usr/local/lib/python3.9/dist-packages/matplotlib/__init__.py in inner(ax, data, *
args, **kwargs)
   1412     def inner(ax, *args, data=None, **kwargs):
   1413         if data is None:
-> 1414             return func(ax, *map(sanitize_sequence, args), **kwargs)
   1415
   1416         bound = new_sig.bind(ax, *args, **kwargs)

/usr/local/lib/python3.9/dist-packages/matplotlib/axes/_axes.py in hist(self, x, b
ins, range, density, weights, cumulative, bottom, histtype, align, orientation, rw
idth, log, color, label, stacked, **kwargs)
   6639             # this will automatically overwrite bins,
   6640             # so that each histogram uses the same bins
-> 6641             m, bins = np.histogram(x[i], bins, weights=w[i], **hist_kwarg
s)
   6642             tops.append(m)
   6643         tops = np.array(tops, float)  # causes problems later if it's an i
nt

/usr/local/lib/python3.9/dist-packages/numpy/core/overrides.py in histogram(*args,
**kwargs)

/usr/local/lib/python3.9/dist-packages/numpy/lib/histograms.py in histogram(a, bin
s, range, normed, weights, density)
    791     a, weights = _ravel_and_check_weights(a, weights)
    792
--> 793     bin_edges, uniform_bins = _get_bin_edges(a, bins, range, weights)
    794
    795     # Histogram is an integer or a float array depending on the weights.

/usr/local/lib/python3.9/dist-packages/numpy/lib/histograms.py in _get_bin_edges
(a, bins, range, weights)
    388         # this will replace it with the number of bins calculated
    389         if bin_name not in _hist_bin_selectors:
--> 390             raise ValueError(
    391                 "{!r} is not a valid estimator for `bins`".format(bin_nam
e))
    392         if weights is not None:

ValueError: 'Age' is not a valid estimator for `bins`
```



In [ ]:
```
df.isnull().sum
```

Out[ ]:
```
<bound method NDFrame._add_numeric_operations.<locals>.sum of      PassengerId  Su
rvived  Pclass   Name    Sex    Age  SibSp  Parch  Ticket  \
0          False     False   False  False  False  False  False  False   False
1          False     False   False  False  False  False  False  False   False
2          False     False   False  False  False  False  False  False   False
3          False     False   False  False  False  False  False  False   False
4          False     False   False  False  False  False  False  False   False
..           ...       ...     ...    ...    ...    ...    ...    ...     ...
886        False     False   False  False  False  False  False  False   False
887        False     False   False  False  False  False  False  False   False
888        False     False   False  False  False   True  False  False   False
889        False     False   False  False  False  False  False  False   False
890        False     False   False  False  False  False  False  False   False

      Fare  Cabin  Embarked
0    False   True     False
1    False  False     False
2    False   True     False
3    False  False     False
4    False   True     False
..     ...    ...       ...
886  False   True     False
887  False  False     False
888  False   True     False
889  False  False     False
890  False   True     False

[891 rows x 12 columns]>
```