

assignment_1.py

May 11, 2022

0.1 importing python dependencies

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

1 Find Dataset

1.1 About Dataset

The food server of a restaurant recorded data about the tips given to the waiters for serving the food. The data recorded by the food server is as follows:

- total_bill: Total bill in dollars including taxes
- tip: Tip given to waiters in dollars
- sex: gender of the person paying the bill
- smoker: whether the person smoked or not
- day: day of the week
- time: lunch or dinner
- size: number of people in a table

```
[3]: #load the dataset
url = "tips.csv"
df = pd.read_csv(url)
df.head()
```

```
[3]:
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

1.1.1 check for null values

```
[4]: print(df.isnull().values.any())
```

False

```
[5]: print(df.describe())
```

	total_bill	tip	size
count	244.000000	244.000000	244.000000
mean	19.785943	2.998279	2.569672
std	8.902412	1.383638	0.951100
min	3.070000	1.000000	1.000000
25%	13.347500	2.000000	2.000000
50%	17.795000	2.900000	2.000000
75%	24.127500	3.562500	3.000000
max	50.810000	10.000000	6.000000

1.1.2 Dimension of the dataframe

```
[6]: print("Size=", df.size)
      print("Shape=", df.shape)
      print("Dimension=", df.ndim)
```

```
Size= 1708
Shape= (244, 7)
Dimension= 2
```

1.1.3 providing variables discriptions

```
[7]: print(df.dtypes)
```

```
total_bill    float64
tip           float64
sex           object
smoker        object
day           object
time          object
size          int64
dtype: object
```

```
[8]: print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
#   Column      Non-Null Count  Dtype
---  -
0   total_bill  244 non-null   float64
1   tip         244 non-null   float64
2   sex         244 non-null   object
3   smoker      244 non-null   object
4   day         244 non-null   object
5   time        244 non-null   object
6   size        244 non-null   int64
dtypes: float64(2), int64(1), object(4)
```

```
memory usage: 13.5+ KB
None
```

1.1.4 type conversion example

```
[9]: df1 = df.copy()
df1["total_bill"] = df1["total_bill"].astype("int64")
df1.head()
```

```
[9]:   total_bill  tip  sex smoker  day  time  size
0         16  1.01 Female    No  Sun  Dinner    2
1         10  1.66  Male    No  Sun  Dinner    3
2         21  3.50  Male    No  Sun  Dinner    3
3         23  3.31  Male    No  Sun  Dinner    2
4         24  3.61 Female    No  Sun  Dinner    4
```

1.1.5 Turn categorical variables into quantitative variables in Python.

```
[10]: df2=df.copy()
df2["female"]=1*(df2["sex"]=="Female")
df2["male"]=1*(df2["sex"]=="Male")
print(df2["male"])
```

```
0      0
1      1
2      1
3      1
4      0
..
239    1
240    0
241    1
242    1
243    0
Name: male, Length: 244, dtype: int64
```

```
[11]: df2.head()
```

```
[11]:   total_bill  tip  sex smoker  day  time  size  female  male
0        16.99  1.01 Female    No  Sun  Dinner    2        1    0
1        10.34  1.66  Male    No  Sun  Dinner    3        0    1
2        21.01  3.50  Male    No  Sun  Dinner    3        0    1
3        23.68  3.31  Male    No  Sun  Dinner    2        0    1
4        24.59  3.61 Female    No  Sun  Dinner    4        1    0
```