Observing and describing data

November 6, 2020

1 Observing and describing data

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

[2]:
```
diamonds_url = "https://raw.githubusercontent.com/TrainingByPackt/
↪Interactive-Data-Visualization-with-Python/master/datasets/diamonds.csv"
```

[3]:
```
diamonds_df = pd.read_csv(diamonds_url)

diamonds_df = sns.load_dataset('diamonds')
```

[4]:
```
diamonds_df.head()
```

[4]:

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

[5]:
```
diamonds_df.shape
```

[5]:
```
(53940, 10)
```

[6]:
```
diamonds_df.describe()
```

[6]:

	carat	depth	table	price	x	\
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	
mean	0.797940	61.749405	57.457184	3932.799722	5.731157	
std	0.474011	1.432621	2.234491	3989.439738	1.121761	
min	0.200000	43.000000	43.000000	326.000000	0.000000	
25%	0.400000	61.000000	56.000000	950.000000	4.710000	
50%	0.700000	61.800000	57.000000	2401.000000	5.700000	
75%	1.040000	62.500000	59.000000	5324.250000	6.540000	
max	5.010000	79.000000	95.000000	18823.000000	10.740000	

	y	z

```
count  53940.000000  53940.000000
mean       5.734526      3.538734
std        1.142135      0.705699
min        0.000000      0.000000
25%        4.720000      2.910000
50%        5.710000      3.530000
75%        6.540000      4.040000
max       58.900000     31.800000
```

[7]:
```
diamonds_df.describe(include=object)
```

[7]:
```
          cut  color clarity
count   53940  53940   53940
unique      5      7       8
top     Ideal      G     SI1
freq    21551  11292   13065
```

[8]:
```
diamonds_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53940 entries, 0 to 53939
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    53940 non-null  float64
 1   cut      53940 non-null  object
 2   color    53940 non-null  object
 3   clarity  53940 non-null  object
 4   depth    53940 non-null  float64
 5   table    53940 non-null  float64
 6   price    53940 non-null  int64
 7   x        53940 non-null  float64
 8   y        53940 non-null  float64
 9   z        53940 non-null  float64
dtypes: float64(6), int64(1), object(3)
memory usage: 4.1+ MB
```

2 Selecting columns from dataframes

[9]:
```
diamonds_low_df = diamonds_df.loc[diamonds_df['cut']=='Ideal']
```

[10]:
```
diamonds_low_df.head()
```

[10]:

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
11	0.23	Ideal	J	VS1	62.8	56.0	340	3.93	3.90	2.46
13	0.31	Ideal	J	SI2	62.2	54.0	344	4.35	4.37	2.71

```
16   0.30  Ideal     I     SI2   62.0   54.0    348  4.31  4.34  2.68
39   0.33  Ideal     I     SI2   61.8   55.0    403  4.49  4.51  2.78
```

[ ]: