DB0201EN-Week3-1-4-Analyzing-v5-py

May 18, 2020

[1]:
```
%load_ext sql
```

[2]:
```
# %sql ibm_db_sa://my-username:my-password@my-hostname:my-port/my-db-name
%sql ibm_db_sa://
```

[2]:
```
'Connected: xpr61119@BLUDB'
```

[4]:
```
import pandas
chicago_socioeconomic_data = pandas.read_csv('https://data.cityofchicago.org/
↪resource/jcxq-k9xf.csv')
# chicago_socioeconomic_data
%sql PERSIST chicago_socioeconomic_data
```

```
 * ibm_db_sa://xpr61119:***@dashdb-txn-sbox-yp-lon02-01.services.eu-
gb.bluemix.net:50000/BLUDB
```

[4]:
```
'Persisted chicago_socioeconomic_data'
```

[5]:
```
%sql SELECT * FROM chicago_socioeconomic_data limit 5;
```

```
 * ibm_db_sa://xpr61119:***@dashdb-txn-sbox-yp-lon02-01.services.eu-
gb.bluemix.net:50000/BLUDB
Done.
```

[5]:
```
[(0, 1.0, 'Rogers Park', 7.7, 23.6, 8.7, 18.2, 27.5, 23939, 39.0),
 (1, 2.0, 'West Ridge', 7.8, 17.2, 8.8, 20.8, 38.5, 23040, 46.0),
 (2, 3.0, 'Uptown', 3.8, 24.0, 8.9, 11.8, 22.2, 35787, 20.0),
 (3, 4.0, 'Lincoln Square', 3.4, 10.9, 8.2, 13.4, 25.5, 37524, 17.0),
 (4, 5.0, 'North Center', 0.3, 7.5, 5.2, 4.5, 26.2, 57123, 6.0)]
```

0.1 Problems

0.1.1 Problem 1

How many rows are in the dataset?

[9]:
```
%sql select count(*) from chicago_socioeconomic_data;
```

```
 * ibm_db_sa://xpr61119:***@dashdb-txn-sbox-yp-lon02-01.services.eu-
gb.bluemix.net:50000/BLUDB
Done.
```

[9]:
```
[(Decimal('78'),)]
```

0.1.2 Problem 2

How many community areas in Chicago have a hardship index greater than 50.0?

[13]:
```
%sql SELECT COUNT(*) FROM chicago_socioeconomic_data WHERE hardship_index > 50.
↪0;
```

```
 * ibm_db_sa://xpr61119:***@dashdb-txn-sbox-yp-lon02-01.services.eu-
gb.bluemix.net:50000/BLUDB
Done.
```

[13]:
```
[(Decimal('38'),)]
```

[14]:
```
%sql SELECT MAX(hardship_index) FROM chicago_socioeconomic_data
```

```
 * ibm_db_sa://xpr61119:***@dashdb-txn-sbox-yp-lon02-01.services.eu-
gb.bluemix.net:50000/BLUDB
Done.
```

[14]:
```
[(98.0,)]
```

0.1.3 Problem 4

Which community area which has the highest hardship index?

[15]:
```
%sql SELECT community_area_name FROM chicago_socioeconomic_data WHERE␣
↪hardship_index = (SELECT MAX(hardship_index) FROM␣
↪chicago_socioeconomic_data);
```

```
 * ibm_db_sa://xpr61119:***@dashdb-txn-sbox-yp-lon02-01.services.eu-
gb.bluemix.net:50000/BLUDB
Done.
```

[15]:
```
[('Riverdale',)]
```

0.1.4 Problem 5

Which Chicago community areas have per-capita incomes greater than $60,000?

[18]:
```
%sql SELECT community_area_name FROM chicago_socioeconomic_data WHERE␣
↪per_capita_income_ > 60000;
```

```
 * ibm_db_sa://xpr61119:***@dashdb-txn-sbox-yp-lon02-01.services.eu-
gb.bluemix.net:50000/BLUDB
Done.
```

[18]:
```
[('Lake View',), ('Lincoln Park',), ('Near North Side',), ('Loop',)]
```

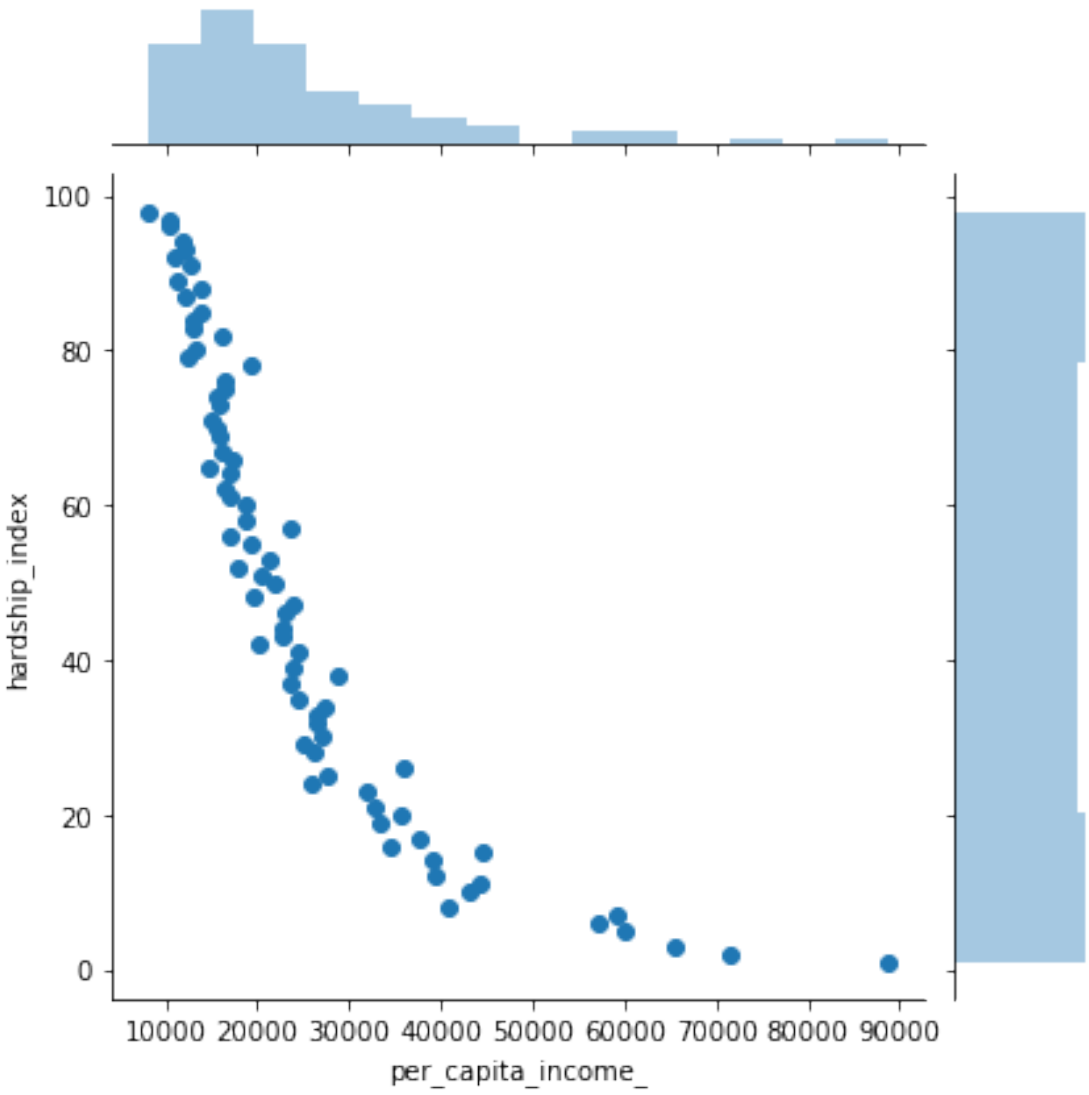
0.1.5 Problem 6

Create a scatter plot using the variables `per_capita_income_` and `hardship_index`. Explain the correlation between the two variables.

[20]:
```
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

income_vs_hardship = %sql SELECT per_capita_income_, hardship_index FROM␣
 ↪chicago_socioeconomic_data;
plot = sns.jointplot(x='per_capita_income_',y='hardship_index',␣
 ↪data=income_vs_hardship.DataFrame())
```

```
 * ibm_db_sa://xpr61119:***@dashdb-txn-sbox-yp-lon02-01.services.eu-
gb.bluemix.net:50000/BLUDB
Done.
```


hardship_index
per_capita_income_
0
20
40
60
80
100
10000
20000
30000
40000
50000
60000
70000
80000
90000