assignment 10 (own)

March 8, 2022

1 Bivariate analysis using Pearson correlation and scatter plot of a combination of 2 columns on Melbourne dataset

For the remaining excersices where numerical columns are needed, I will use the Melbourne Housing Market dataset
https://www.kaggle.com/anthonypino/melbourne-housing-market?select=MELBOURNE_HOUSE_PRICES_LI

Some Key Details

Suburb: Suburb

Address: Address

Rooms: Number of rooms

Price: Price in Australian dollars

Method:
S - property sold;
SP - property sold prior;
PI - property passed in;
PN - sold prior not disclosed;
SN - sold not disclosed;
NB - no bid;
VB - vendor bid;
W - withdrawn prior to auction;
SA - sold after auction;
SS - sold after auction price not disclosed.
N/A - price or highest bid not available.

Type:
br - bedroom(s);
h - house,cottage,villa, semi,terrace;
u - unit, duplex;
t - townhouse;
dev site - development site;
o res - other residential.

SellerG: Real Estate Agent

Distance: Distance from CBD (Central Business District) in Kilometres

Date: Date sold

Regionname: General Region (West, North West, North, North east …etc)

Propertycount: Number of properties that exist in the suburb.

Bedroom2 : Scraped # of Bedrooms (from different source)

Bathroom: Number of Bathrooms

Car: Number of carspots

Landsize: Land Size in Metres

BuildingArea: Building Size in Metres

YearBuilt: Year the house was built

CouncilArea: Governing council for the area

Lattitude: Self explanitory

Longtitude: Self explanitory

[2]:
```
import pandas as pd
import seaborn as sns
```

[3]:
```
df = pd.read_csv('melbourne_housing_prices.csv', sep=',')
df.head()
```

[3]:

	Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Postcode	Regionname	Propertycount	Distance	CouncilArea
0	Abbotsford	49 Lithgow St	3	h	1490000.0	S	Jellis	1/04/2017	3067	Northern Metropolitan	4019	3.0	Yarra City Council
1	Abbotsford	59A Turner St	3	h	1220000.0	S	Marshall	1/04/2017	3067	Northern Metropolitan	4019	3.0	Yarra City Council
2	Abbotsford	119B Yarra St	3	h	1420000.0	S	Nelson	1/04/2017	3067	Northern Metropolitan	4019	3.0	Yarra City Council
3	Aberfeldie	68 Vida St	3	h	1515000.0	S	Barry	1/04/2017	3040	Western Metropolitan	1543	7.5	Moonee Valley City Council
4	Airport West	92 Clydesdale Rd	2	h	670000.0	S	Nelson	1/04/2017	3042	Western Metropolitan	3464	10.4	Moonee Valley City Council

[9]:
```
houseCorrelations = df.corr(method='pearson')
```

```
houseCorrelations.style.background_gradient(cmap='coolwarm', axis=None).
 ↪set_precision(2)
```

```
C:\Users\Stijn\AppData\Local\Temp/ipykernel_13424/2374652365.py:2:
FutureWarning: this method is deprecated in favour of
`Styler.format(precision=..)`
  houseCorrelations.style.background_gradient(cmap='coolwarm',
axis=None).set_precision(2)
```

[9]: <pandas.io.formats.style.Styler at 0x27f52100670>

I expected distance to the business district and price to correlate negatively, I did expect a stronger correlation

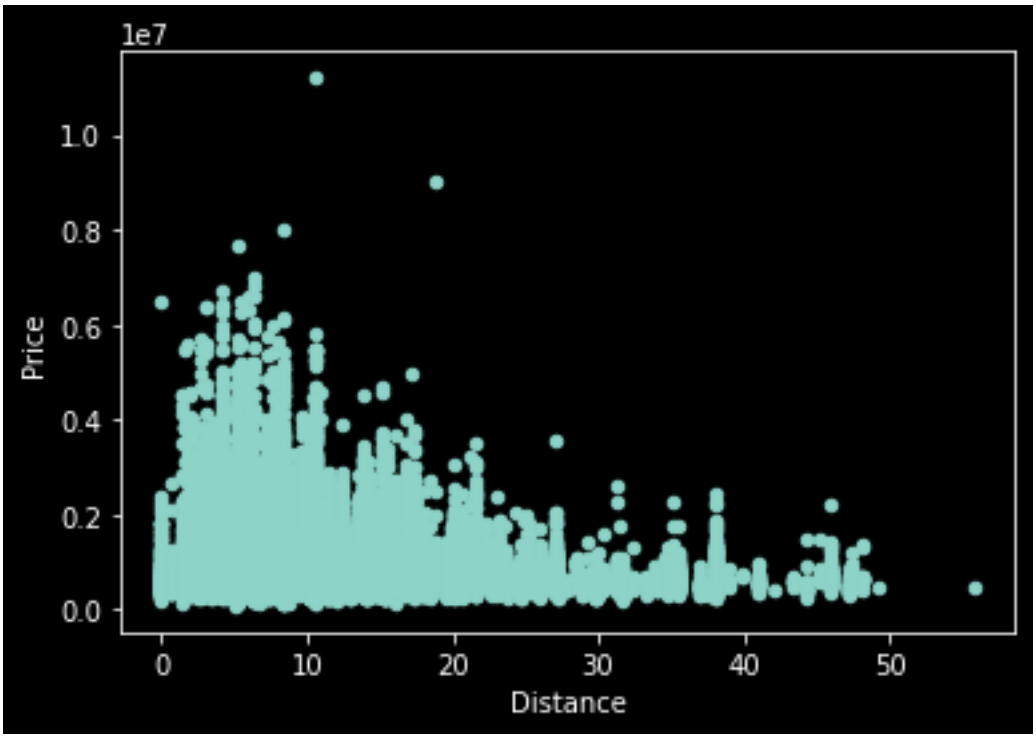
[5]:
```
correlation = df[['Price','Distance']]
correlation.corr()
```

[5]:

	Price	Distance
Price	1.000000	-0.253668
Distance	-0.253668	1.000000

[7]:
```
df.plot(kind='scatter', x='Distance',y='Price')
```

[7]: <AxesSubplot:xlabel='Distance', ylabel='Price'>



The scatter plot shows the negative correlation very well.