assignment 12 (own)

March 8, 2022

1 Bivariate analysis on 3 numerical columns with 1 categorical columns

[13]:
```
import pandas as pd
```

[14]:
```
import seaborn as sns
```

[15]:
```
df = pd.read_csv('melbourne_housing_prices.csv', sep=',')
```

[16]:
```
df.head()
```

[16]:
```
          Suburb           Address  Rooms Type      Price Method   SellerG  \
0    Abbotsford     49 Lithgow St      3    h  1490000.0      S    Jellis
1    Abbotsford     59A Turner St      3    h  1220000.0      S  Marshall
2    Abbotsford     119B Yarra St      3    h  1420000.0      S    Nelson
3    Aberfeldie        68 Vida St      3    h  1515000.0      S     Barry
4  Airport West  92 Clydesdale Rd      2    h   670000.0      S    Nelson

        Date  Postcode             Regionname  Propertycount  Distance  \
0  1/04/2017      3067  Northern Metropolitan           4019       3.0
1  1/04/2017      3067  Northern Metropolitan           4019       3.0
2  1/04/2017      3067  Northern Metropolitan           4019       3.0
3  1/04/2017      3040   Western Metropolitan           1543       7.5
4  1/04/2017      3042   Western Metropolitan           3464      10.4

                  CouncilArea
0          Yarra City Council
1          Yarra City Council
2          Yarra City Council
3  Moonee Valley City Council
4  Moonee Valley City Council
```

I expect there to be a difference of the amount of rooms and price per region, as richer regions will have bigger houses (with more rooms) and higher prices and vice versa.

[17]:
```
df.groupby(by='Regionname').mean()
```

[17]:
```
                               Rooms         Price     Postcode  \
Regionname
Eastern Metropolitan        3.436129  1.071837e+06  3123.375337
Eastern Victoria            3.560284  6.893688e+05  3503.820922
Northern Metropolitan       2.922591  8.237198e+05  3097.344318
Northern Victoria           3.559353  6.430822e+05  3458.339928
South-Eastern Metropolitan  3.281082  8.462352e+05  3222.596124
Southern Metropolitan       2.984168  1.373222e+06  3157.935873
Western Metropolitan        3.158658  8.005163e+05  3038.528975
Western Victoria            3.260504  4.120261e+05  3337.806723

                            Propertycount   Distance
Regionname
Eastern Metropolitan          6943.980954  15.396749
Eastern Victoria              9147.015957  33.749291
Northern Metropolitan         9386.327275  10.583338
Northern Victoria             4374.142086  33.387410
South-Eastern Metropolitan    6993.424597  24.457751
Southern Metropolitan         7387.549177   8.910097
Western Metropolitan          6458.604421  11.337399
Western Victoria              4018.415966  31.125630
```

There seem to be significant differences in the amount of rooms and price per region, as expected.

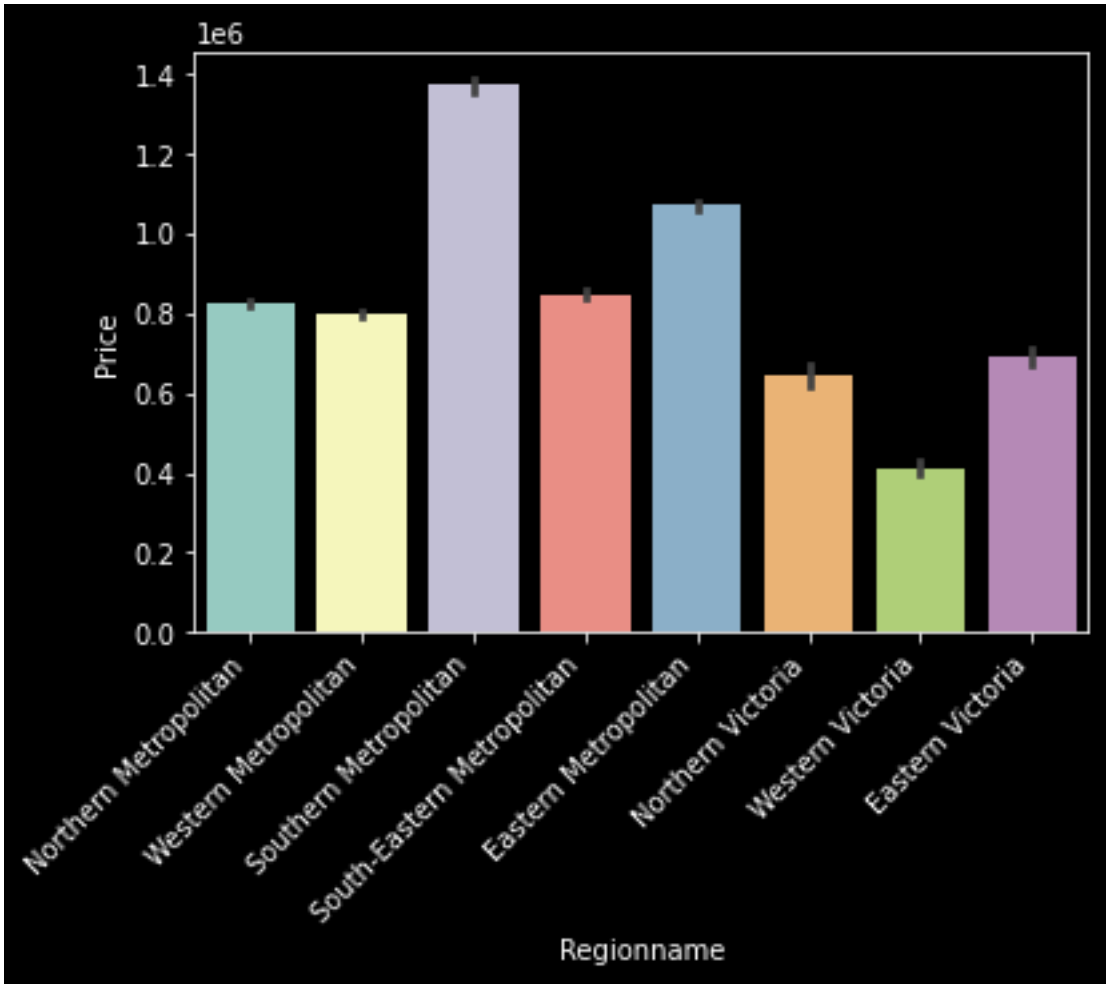
2 Average price per region

Let's see if there are significant price differences

Numbers are in millions (le6 = 6 zeros)

[18]:
```
plot = sns.barplot(y="Price",x="Regionname",data=df)
plot.set_xticklabels(plot.get_xticklabels(), rotation=45,␣
↪horizontalalignment='right')
```

[18]:
```
[Text(0, 0, 'Northern Metropolitan'),
 Text(1, 0, 'Western Metropolitan'),
 Text(2, 0, 'Southern Metropolitan'),
 Text(3, 0, 'South-Eastern Metropolitan'),
 Text(4, 0, 'Eastern Metropolitan'),
 Text(5, 0, 'Northern Victoria'),
 Text(6, 0, 'Western Victoria'),
 Text(7, 0, 'Eastern Victoria')]
```



There are large statistically significant differences. For example, Western Victoria is 3 times less expensive than Southern Metropolitan

3 Amount of average rooms per home per region

Does the higher price also mean that the homes are bigger?

[19]:
```
df.groupby(by='Regionname').mean()
```

[19]:

Regionname	Rooms	Price	Postcode
Eastern Metropolitan	3.436129	1.071837e+06	3123.375337
Eastern Victoria	3.560284	6.893688e+05	3503.820922
Northern Metropolitan	2.922591	8.237198e+05	3097.344318
Northern Victoria	3.559353	6.430822e+05	3458.339928
South-Eastern Metropolitan	3.281082	8.462352e+05	3222.596124
Southern Metropolitan	2.984168	1.373222e+06	3157.935873

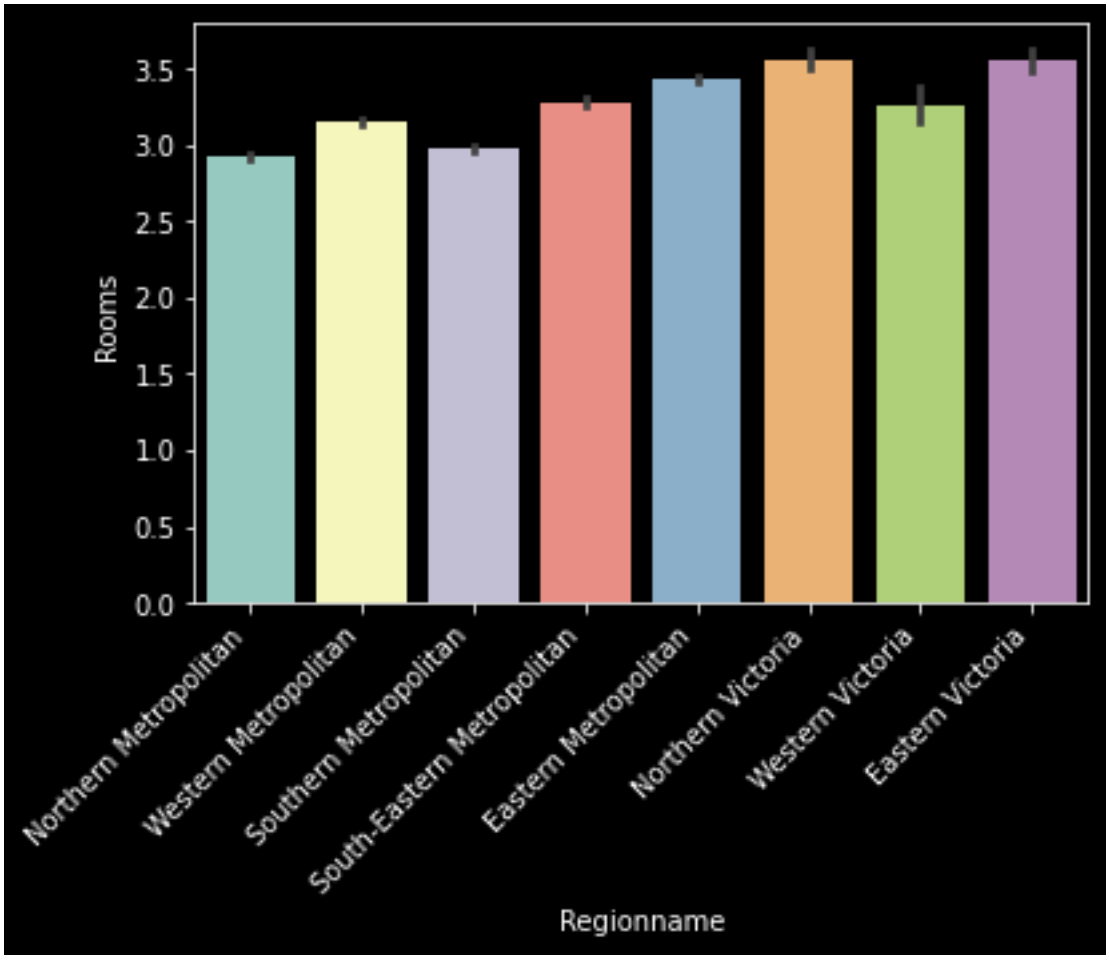
```
Western Metropolitan        3.158658  8.005163e+05  3038.528975
Western Victoria            3.260504  4.120261e+05  3337.806723

                           Propertycount   Distance
Regionname
Eastern Metropolitan         6943.980954  15.396749
Eastern Victoria             9147.015957  33.749291
Northern Metropolitan        9386.327275  10.583338
Northern Victoria            4374.142086  33.387410
South-Eastern Metropolitan   6993.424597  24.457751
Southern Metropolitan        7387.549177   8.910097
Western Metropolitan         6458.604421  11.337399
Western Victoria             4018.415966  31.125630
```

It seems that there aren't significant differences and there are also to many sellers to put them in a barplot

[20]:
```
plot = sns.barplot(y="Rooms",x="Regionname",data=df)
plot.set_xticklabels(plot.get_xticklabels(), rotation=45,␣
↪horizontalalignment='right')
```

[20]:
```
[Text(0, 0, 'Northern Metropolitan'),
 Text(1, 0, 'Western Metropolitan'),
 Text(2, 0, 'Southern Metropolitan'),
 Text(3, 0, 'South-Eastern Metropolitan'),
 Text(4, 0, 'Eastern Metropolitan'),
 Text(5, 0, 'Northern Victoria'),
 Text(6, 0, 'Western Victoria'),
 Text(7, 0, 'Eastern Victoria')]
```



It seems that houses everywhere in Melbourne that are being sold are about the same size: 3 rooms. That's quite small. Only Victoria jumps out as having slightly larger homes with homes nearing 3.5 rooms on average.

4 Amount of homes sold per region

Let's now see where most homes are being sold.

[21]:
```
df.groupby(by='Regionname').mean()
```

[21]:

Regionname	Rooms	Price	Postcode
Eastern Metropolitan	3.436129	1.071837e+06	3123.375337
Eastern Victoria	3.560284	6.893688e+05	3503.820922
Northern Metropolitan	2.922591	8.237198e+05	3097.344318
Northern Victoria	3.559353	6.430822e+05	3458.339928
South-Eastern Metropolitan	3.281082	8.462352e+05	3222.596124
Southern Metropolitan	2.984168	1.373222e+06	3157.935873

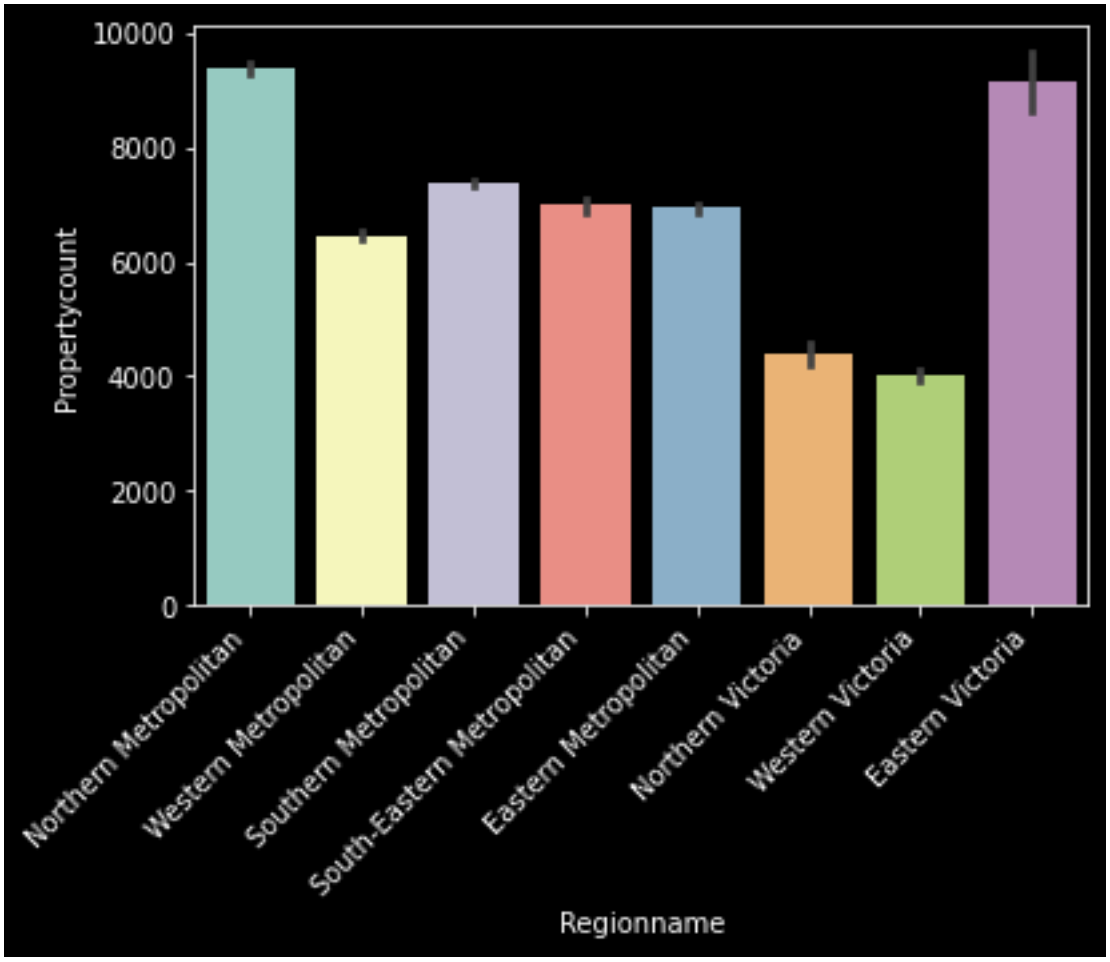
```
Western Metropolitan        3.158658  8.005163e+05  3038.528975
Western Victoria            3.260504  4.120261e+05  3337.806723

                           Propertycount   Distance
Regionname
Eastern Metropolitan         6943.980954  15.396749
Eastern Victoria             9147.015957  33.749291
Northern Metropolitan        9386.327275  10.583338
Northern Victoria            4374.142086  33.387410
South-Eastern Metropolitan   6993.424597  24.457751
Southern Metropolitan        7387.549177   8.910097
Western Metropolitan         6458.604421  11.337399
Western Victoria             4018.415966  31.125630
```

Seems there are large differences. Let's see in detail.

[22]:
```
plot = sns.barplot(y="Propertycount",x="Regionname",data=df)
plot.set_xticklabels(plot.get_xticklabels(), rotation=45,␣
↪horizontalalignment='right')
```

[22]:
```
[Text(0, 0, 'Northern Metropolitan'),
 Text(1, 0, 'Western Metropolitan'),
 Text(2, 0, 'Southern Metropolitan'),
 Text(3, 0, 'South-Eastern Metropolitan'),
 Text(4, 0, 'Eastern Metropolitan'),
 Text(5, 0, 'Northern Victoria'),
 Text(6, 0, 'Western Victoria'),
 Text(7, 0, 'Eastern Victoria')]
```



There are indeed very large differences between regions. Eastern Victoria and Northern Metropolitan are especially hot markets.