014-size-or-location

April 25, 2022

Location or Size: What Influences House Prices in Mexico?

[1]:
```
import matplotlib.pyplot as plt
import pandas as pd
from IPython.display import VimeoVideo
```

You've wrangled the data, you've gained an understanding of its basic characteristics in your EDA, and now it's time to ask some research questions.

1 Import Data

Task 1.4.1: Read the CSV file that you created in the last notebook ("../small-data/mexico-real-estate-clean.csv") into a DataFrame named df. Be sure to check that all your columns are the correct data type before you go to the next task.

- What's a DataFrame?
- What's a CSV file?
- Read a CSV file into a DataFrame using pandas.

[2]:
```
df = pd.read_csv('data/mexico-real-estate-clean.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1736 entries, 0 to 1735
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   property_type  1736 non-null   object
 1   state          1736 non-null   object
 2   lat            1736 non-null   float64
 3   lon            1736 non-null   float64
 4   area_m2        1736 non-null   float64
 5   price_usd      1736 non-null   float64
dtypes: float64(4), object(2)
memory usage: 81.5+ KB
```

2 Research Question 1

Which state has the most expensive real estate market?

Do housing prices vary by state? If so, which are the most expensive states for purchasing a home? During our exploratory data analysis, we used descriptive statistics like mean and median to get an idea of the "typical" house price in Mexico. Now, we need to break that calculation down by state and visualize the results.

We know in which state each house is located thanks to the `"state"` column. The next step is to divide our dataset into groups (one per state) and calculate the mean house price for each group.

```
[3]: VimeoVideo("656378731", h="8daa35d1e8", width=600)
```

```
[3]: <IPython.lib.display.VimeoVideo at 0x7fbf16e1a340>
```

Task 1.4.2: Use the `groupby` method to create a Series named `mean_price_by_state`, where the index contains each state in the dataset and the values correspond to the mean house price for that state. Make sure your Series is sorted from highest to lowest mean price.

- What's a Series?
- Aggregate data using the `groupby` method in pandas.

```
[8]: mean_price_by_state = df.groupby('state')['price_usd'].mean().
      ↪sort_values(ascending = False)
     mean_price_by_state
```

```
[8]: state
     Querétaro                          133955.913281
     Guanajuato                         133277.965833
     Nuevo León                         129221.985663
     Distrito Federal                   128347.267426
     Quintana Roo                       128065.416053
     Chihuahua                          127073.852000
     Jalisco                            123386.472167
     Estado de México                   122723.490503
     Campeche                           121734.633333
     Puebla                             121732.974000
     Guerrero                           119854.276122
     Sonora                             114547.883333
     Morelos                            112697.295625
     Aguascalientes                     110543.888000
     Baja California Sur                109069.339333
     Yucatán                            108580.388596
     Chiapas                            104342.313273
     Veracruz de Ignacio de la Llave     96928.125128
     Hidalgo                             94012.326471
     Sinaloa                             93922.152308
     Tamaulipas                          93713.385833
     San Luis Potosí                     92435.540364
     Nayarit                             87378.605556
     Tabasco                             82763.587000
     Durango                             78034.511429
```

```
Zacatecas                                    76395.400000
Tlaxcala                                     72921.818333
Colima                                       65786.646000
Baja California                              63152.431724
Oaxaca                                       59681.585000
Name: price_usd, dtype: float64
```

[9]:

```
VimeoVideo("656378435", h="b3765f3339", width=600)
```

[9]:

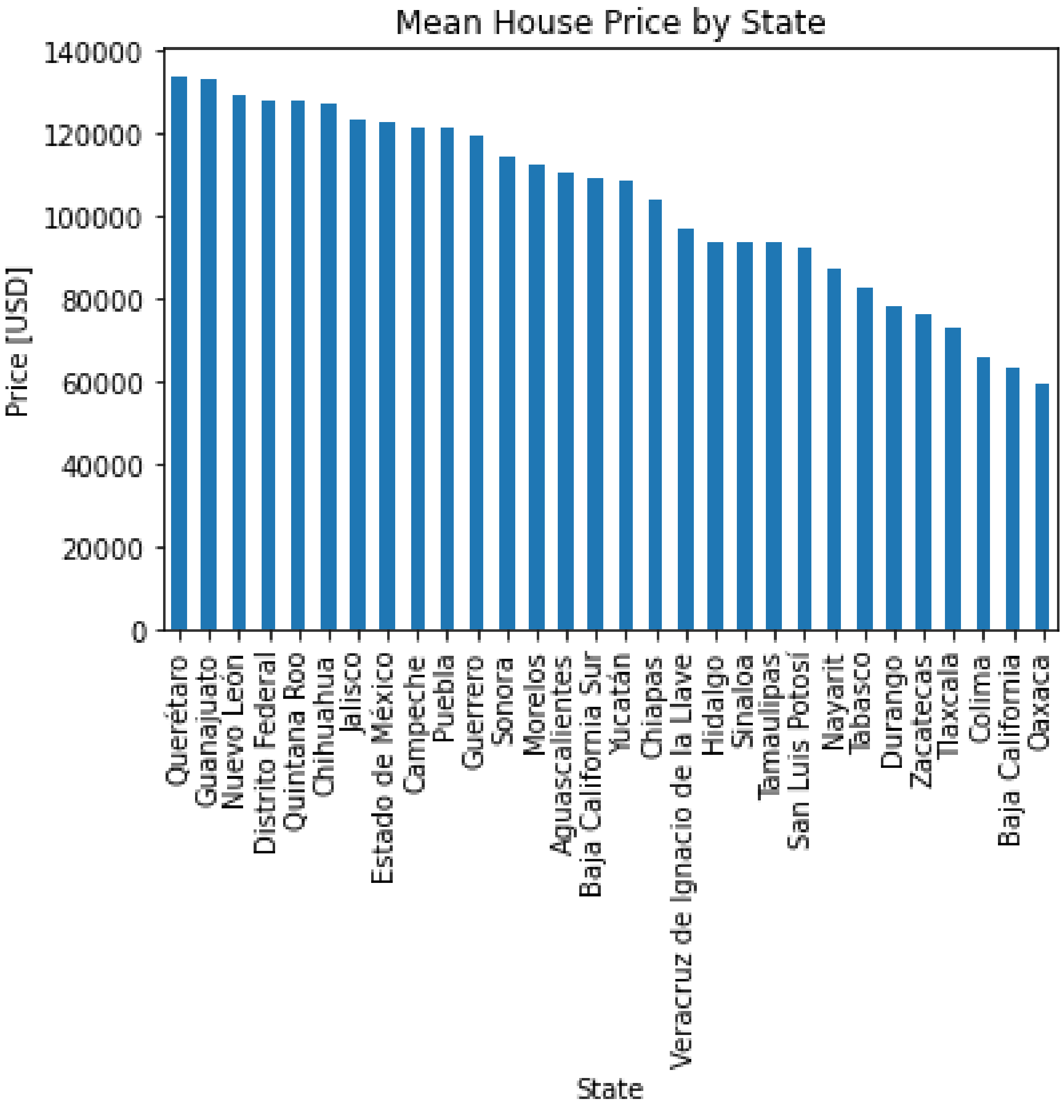
```
<IPython.lib.display.VimeoVideo at 0x7fbf16e1ac40>
```

Task 1.4.3: Use `mean_price_by_state` to create a bar chart of your results. Make sure the states are sorted from the highest to lowest mean, that you label the x-axis as `"State"` and the y-axis as `"Mean Price [USD]"`, and give the chart the title `"Mean House Price by State"`.

- Create a bar chart using pandas.

[12]:

```
mean_price_by_state.plot(
    kind = 'bar',
    xlabel = 'State',
    ylabel = 'Price [USD]',
    title = 'Mean House Price by State'
)
plt.show()
```



It seems odd that Querétaro would be the most expensive real estate market in Mexico when, according to recent GDP numbers, it's not in the top 10 state economies. With all the variations in house sizes across states, a better metric to look at would be price per m2. In order to do that, we need to create a new column.

[13]:
```
VimeoVideo("656378342", h="2f4da7f7b4", width=600)
```

[13]:
```
<IPython.lib.display.VimeoVideo at 0x7fbf14a67730>
```

Task 1.4.4: Create a new column in `df` called `"price_per_m2"`. This should be the price for each house divided by it's size.

- Create new columns derived from existing columns in a DataFrame using pandas.

[14]:

```
df["price_per_m2"] = df['price_usd'] / df['area_m2']
```

[15]:

```
df.head()
```

[15]:

```
  property_type               state        lat         lon  area_m2  price_usd  \
0         house  Estado de México  19.560181  -99.233528    150.0   67965.56
1         house        Nuevo León  25.688436 -100.198807    186.0   63223.78
2     apartment          Guerrero  16.767704  -99.764383     82.0   84298.37
3     apartment          Guerrero  16.829782  -99.911012    150.0   94308.80
4         house           Yucatán  21.052583  -89.538639    205.0  105191.37

   price_per_m2
0    453.103733
1    339.912796
2   1028.028902
3    628.725333
4    513.128634
```

Let's redo our bar chart from above, but this time with the mean of "price_per_m2" for each state.

[16]:

```
VimeoVideo("656377991", h="c7319b0458", width=600)
```

[16]:

```
<IPython.lib.display.VimeoVideo at 0x7fbf14bc21f0>
```

Task 1.4.5: First, use the groupby method to create a Series where the index contains each state in the dataset and the values correspond to the mean house price per m2 for that state. Then use the Series to create a bar chart of your results. Make sure the states are sorted from the highest to lowest mean, that you label the x-axis as "State" and the y-axis as "Mean Price per M^2[USD]", and give the chart the title "Mean House Price per M^2 by State".

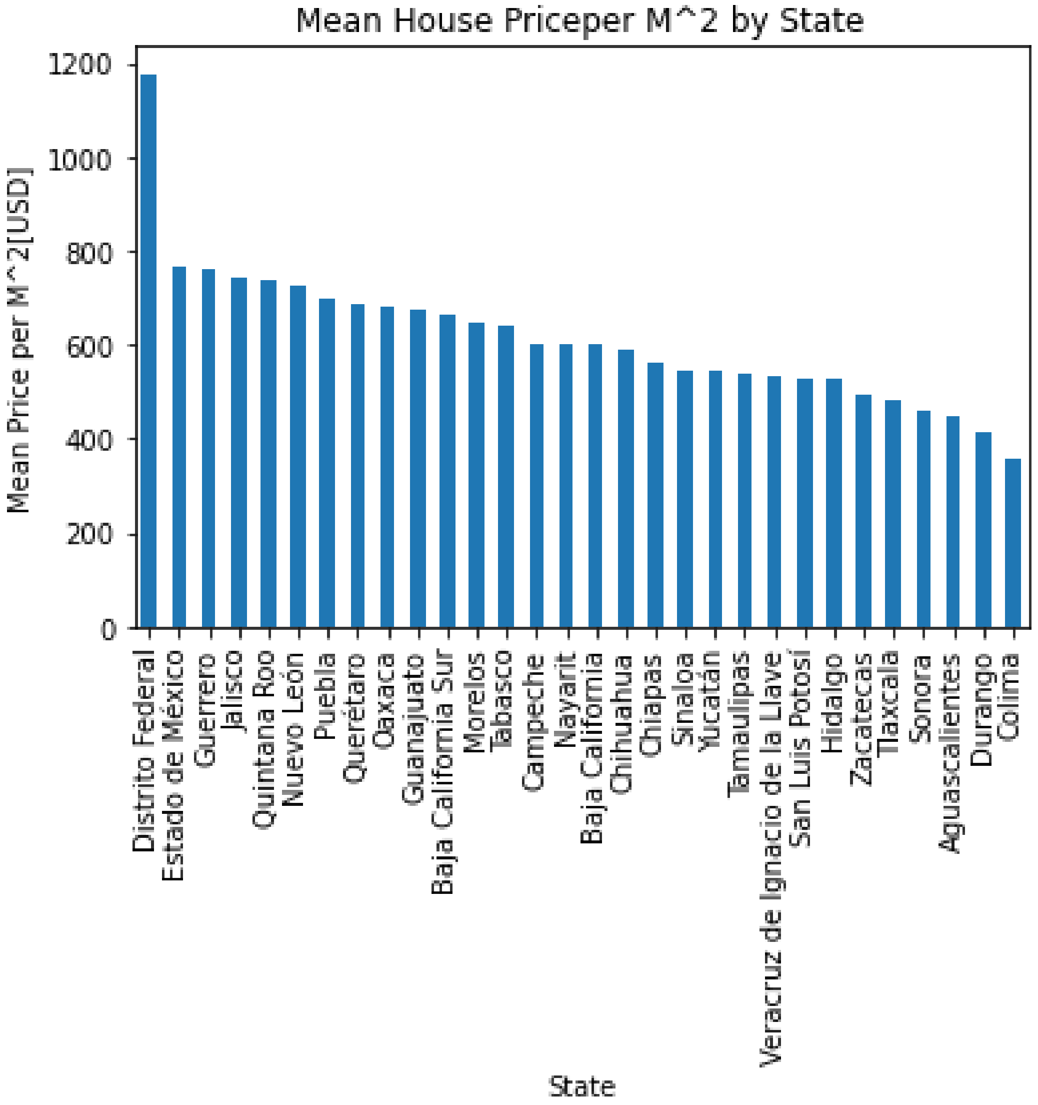
- What's a Series?
- Aggregate data using the groupby method in pandas.
- Create a bar chart using pandas.

[20]:

```
(
df
    .groupby('state')
    ['price_per_m2'].mean()
    .sort_values(ascending = False)
    .plot(
        kind = 'bar',
        xlabel = 'State',
        ylabel = 'Mean Price per M^2[USD]',
        title = 'Mean House Priceper M^2 by State'

    )
)
```

```
plt.show()
```



Now we see that the capital Mexico City (*Distrito Federal*) is by far the most expensive market. Additionally, many of the top 10 states by GDP are also in the top 10 most expensive real estate markets. So it looks like this bar chart is a more accurate reflection of state real estate markets.

3 Research Question 2

Is there a relationship between home size and price?

From our previous question, we know that the location of a home affects its price (especially if it's in Mexico City), but what about home size? Does the size of a house influence price?

A scatter plot can be helpful when evaluating the relationship between two columns because it lets you see if two variables are correlated — in this case, if an increase in home size is associated with an increase in price.

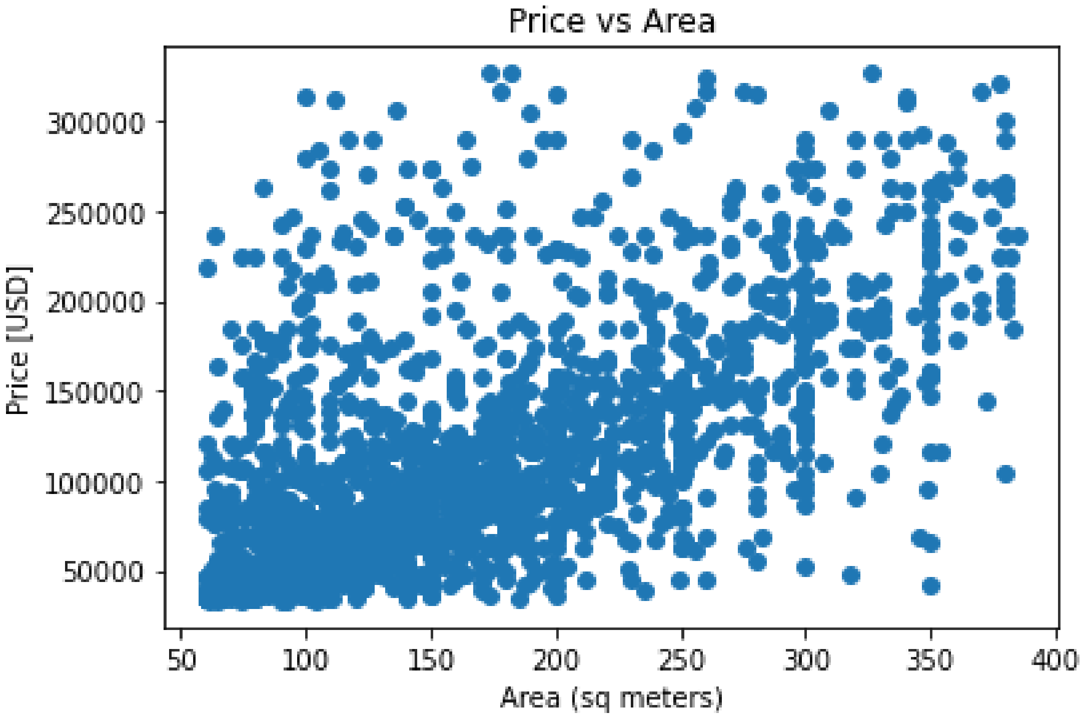
```
[21]: VimeoVideo("656377758", h="62546c7b86", width=600)
```

```
[21]: <IPython.lib.display.VimeoVideo at 0x7fbf148492e0>
```

Task 1.4.6: Create a scatter plot from `df` that represents price as a function of size. In other words, `"area_m2"` should be on the x-axis, and `"price_usd"` should be on the y-axis. Be sure to use expressive axis labels (`"Area [sq meters]"` and `"Price [USD]"`, respectively).

- What's a scatter plot?
- What's correlation?
- Create a scatter plot using Matplotlib.

```
[25]: plt.scatter(x = df['area_m2'], y = df['price_usd'])
      plt.xlabel('Area (sq meters)')
      plt.ylabel('Price [USD]')
      plt.title('Price vs Area');
```



While there's a good amount of variation, there's definitely a positive correlation — in other words, the bigger the house, the higher the price. But how can we quantify this correlation?

```
[26]: VimeoVideo("656377616", h="8d3b060e71", width=600)
```

```
[26]: <IPython.lib.display.VimeoVideo at 0x7fbf14897100>
```

Task 1.4.7: Using the `corr` method, calculate the Pearson correlation coefficient for `"area_m2"` and `"price_usd"`.

- What's a correlation coefficient?
- Calculate the correlation coefficient for two Series using pandas.

```
[27]: p_correlation = df['area_m2'].corr(df['price_usd'])
      print(p_correlation)
```

```
0.5855182453232062
```

The correlation coefficient is over 0.5, so there's a moderate relationship house size and price in Mexico. But does this relationship hold true in every state? Let's look at a couple of states, starting with Morelos.

```
[28]: VimeoVideo("656377515", h="d2478d38df", width=600)
```

```
[28]: <IPython.lib.display.VimeoVideo at 0x7fbf147601f0>
```

Task 1.4.8: Create a new DataFrame named `df_morelos`. It should include all the houses from `df` that are in the state of Morelos.

- Subset a DataFrame with a mask using pandas.

```
[29]: df_morelos = df[df['state'] == 'Morelos']
      df_morelos.head()
```

```
[29]:    property_type    state        lat        lon  area_m2  price_usd  \
      6          house  Morelos  18.812605 -98.954826    281.0  151509.56
      9          house  Morelos  18.804197 -98.932816    117.0   63223.78
      18         house  Morelos  18.855343 -99.241142     73.0   36775.16
      49         house  Morelos  18.804197 -98.932816    130.0   65858.10
      55         house  Morelos  18.960244 -99.212962    305.0  227351.46

          price_per_m2
      6     539.179929
      9     540.374188
      18    503.769315
      49    506.600769
      55    745.414623
```

```
[30]: VimeoVideo("656377395", h="bd93b05ff9", width=600)
```

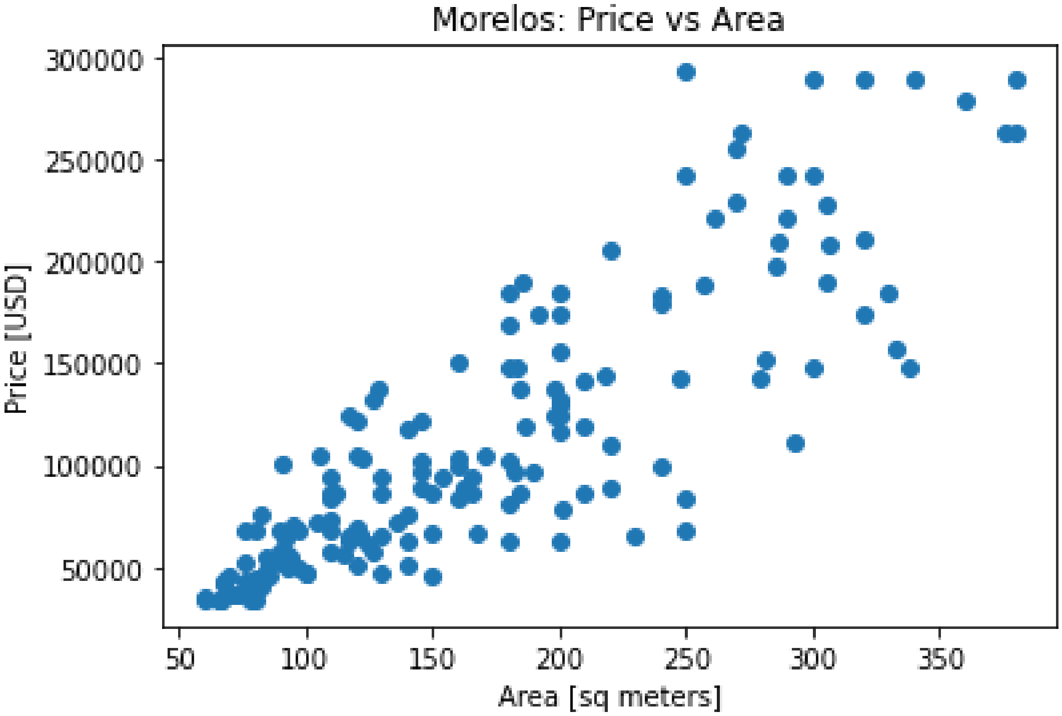
```
[30]: <IPython.lib.display.VimeoVideo at 0x7fbf16e6cfd0>
```

Task 1.4.9: Using `df_morelos`, create a scatter plot that shows price vs area. Make sure to use the same axis labels as your last scatter plot. The title should be `"Morelos: Price vs. Area"`.

- What's a scatter plot?

- Create a scatter plot using Matplotlib.

[31]:
```
plt.scatter(x = df_morelos['area_m2'], y = df_morelos['price_usd'])
plt.xlabel('Area [sq meters]')
plt.ylabel('Price [USD]')
plt.title('Morelos: Price vs Area');
```



Wow! It looks like the correlation is even stronger within Morelos. Let's calculate the correlation coefficient and verify that that's the case.

[32]:
```
VimeoVideo("656377340", h="664cb44291", width=600)
```

[32]:
```
<IPython.lib.display.VimeoVideo at 0x7fbf146456d0>
```

Task 1.4.10: Using the `corr` method, calculate the Pearson correlation coefficient for `"area_m2"` and `"price_usd"` in `df_morelos`.

- What's a correlation coefficient?
- Calculate the correlation coefficient for two Series using pandas.

[33]:
```
p_correlation = df_morelos['area_m2'].corr(df_morelos['price_usd'])
print(p_correlation)
```

```
0.8498077608713708
```

With a correlation coefficient that high, we can say that there's a strong relationship between house size and price in Morelos.

To conclude, let's look at the capital Mexico City (*Distrito Federal*).

```
[34]: VimeoVideo("656376911", h="19666a4c87", width=600)
```

```
[34]: <IPython.lib.display.VimeoVideo at 0x7fbf14645670>
```

Task 1.4.11: First, create a new DataFrame called `df_mexico_city` that includes all the observations from `df` that are part of the *Distrito Federal*. Next, create a scatter plot that shows price vs area. Don't forget to label the x- and y-axis and use the title `"Mexico City: Price vs. Area"`. Finally, calculate the correlation coefficient for `"area_m2"` and `"price_usd"` in `df_mexico_city`.

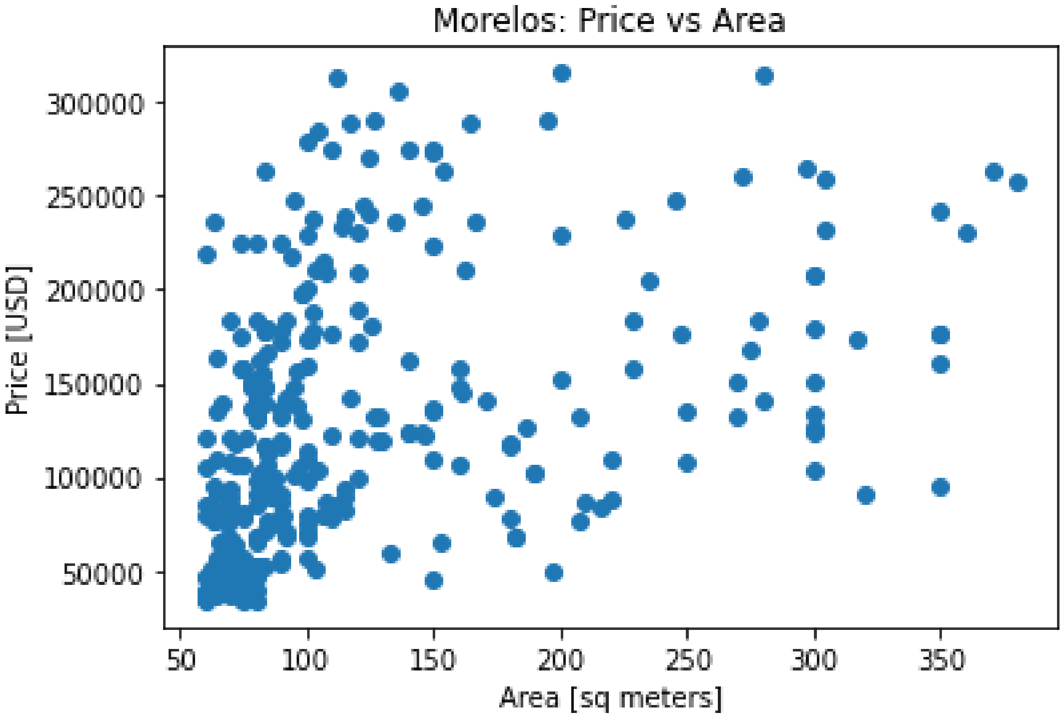
- Calculate the correlation coefficient for two Series using pandas.
- Create a scatter plot using Matplotlib.
- Subset a DataFrame with a mask using pandas.

```
[35]: # Subset `df` to include only observations from `"Distrito Federal"`
df_mexico_city = df[df['state'] == 'Distrito Federal']

# Create a scatter plot price vs area
plt.scatter(x = df_mexico_city['area_m2'], y = df_mexico_city['price_usd'])
plt.xlabel('Area [sq meters]')
plt.ylabel('Price [USD]')
plt.title('Morelos: Price vs Area');

p_correlation = df_mexico_city['area_m2'].corr(df_mexico_city['price_usd'])
print(p_correlation)
```

```
0.41070392130717887
```



Looking at the scatter plot and correlation coefficient, there's see a weak relationship between size and price. How should we interpret this?

One interpretation is that the relationship we see between size and price in many states doesn't hold true in the country's biggest and most economically powerful urban center because there are other factors that have a larger influence on price. In fact, in the next project, we're going to look at another important Latin American city — Buenos Aires, Argentina — and build a model that predicts housing price by taking much more than size into account.