014-size-or-location

May 9, 2022

Location or Size: What Influences House Prices in Mexico?

[3]:
```
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.

[4]:
```
df =  pd.read_csv("data/mexico_real_estate_clean.csv")

df.shape
df.info()

df.head(10)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1159 entries, 0 to 1158
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   property_type  1159 non-null   object 
 1   state          1159 non-null   object 
 2   lat            1159 non-null   float64
 3   lon            1159 non-null   float64
 4   area_m2        1159 non-null   float64
 5   price_usd      1159 non-null   float64
dtypes: float64(4), object(2)
memory usage: 54.5+ KB
```

[4]:

	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
5	house	Querétaro	20.716315	-100.452503	320.0	274034.68
6	house	Morelos	18.812605	-98.954826	281.0	151509.56
7	house	Chiapas	16.769737	-93.088928	140.0	79029.72
8	house	Estado de México	19.305407	-99.646948	235.0	115937.75
9	house	Morelos	18.804197	-98.932816	117.0	63223.78

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.

[5]:
```
VimeoVideo("656378731", h="8daa35d1e8", width=600)
```

[5]:
```
<IPython.lib.display.VimeoVideo at 0x7f0f7dacc1f0>
```

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.

[6]:
```
mean_price_by_state = df.groupby("state")["price_usd"].mean().
↪sort_values(ascending=False)
mean_price_by_state
```

[6]:
```
state
Querétaro                 141521.234079
Guanajuato                138934.256000
Distrito Federal          137502.272525
Chihuahua                 132085.373333
Quintana Roo              130142.436400
Estado de México          124329.215766
Puebla                    123336.021781
Guerrero                  115691.668919
```

```
Nuevo León                          112529.309623
Jalisco                             110828.913415
Sonora                              109995.920000
Yucatán                             109715.606462
Morelos                             108134.703196
Aguascalientes                      104197.078750
Chiapas                             103286.501562
Baja California Sur                  97075.972500
Hidalgo                              95241.989167
Tamaulipas                           94822.754865
Veracruz de Ignacio de la Llave      93383.069125
Sinaloa                              92547.207778
San Luis Potosí                      83595.254872
Durango                              83590.000000
Tlaxcala                             80340.822000
Nayarit                              77654.416250
Zacatecas                            76395.400000
Baja California                      65993.726842
Tabasco                              63247.572857
Colima                               63157.890000
Name: price_usd, dtype: float64
```

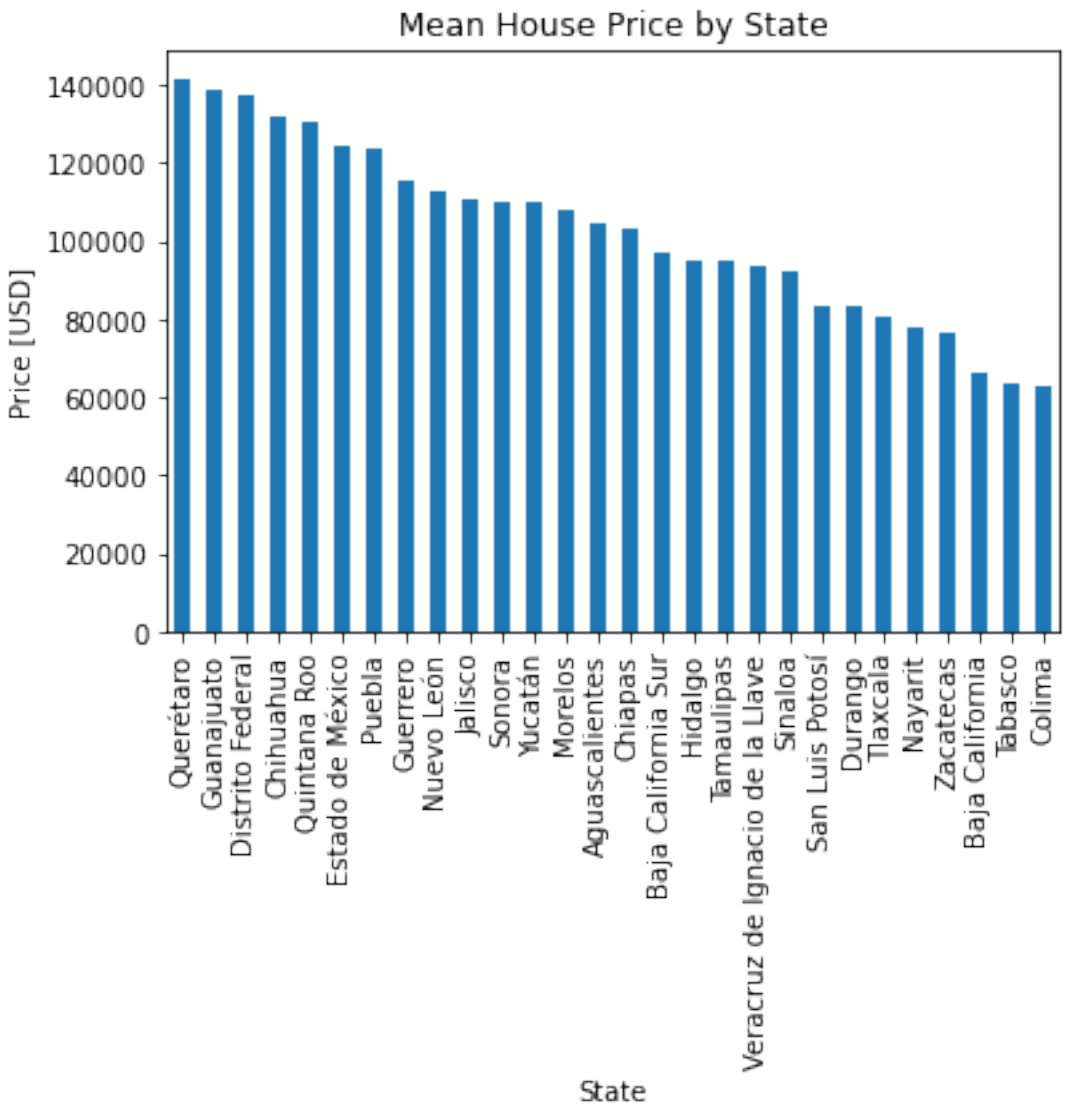
[7]:
```
VimeoVideo("656378435", h="b3765f3339", width=600)
```

[7]:
```
<IPython.lib.display.VimeoVideo at 0x7f0f7de1fdc0>
```

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.

[8]:
```
mean_price_by_state.plot(
    kind="bar",
    xlabel="State",
    ylabel="Price [USD]",
    title="Mean House Price by State"
);
```



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.

[9]:
```
VimeoVideo("656378342", h="2f4da7f7b4", width=600)
```

[9]:
```
<IPython.lib.display.VimeoVideo at 0x7f0f7b0ecdf0>
```

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.

[10]:
```
df["price_per_m2"] = df["price_usd"]/df["area_m2"]
df.head(10)
```

[10]:

	property_type	state	lat	lon	area_m2	price_usd	price_per_m2
0	house	Estado de México	19.560181	-99.233528	150.0	67965.56	453.103733
1	house	Nuevo León	25.688436	-100.198807	186.0	63223.78	339.912796
2	apartment	Guerrero	16.767704	-99.764383	82.0	84298.37	1028.028902
3	apartment	Guerrero	16.829782	-99.911012	150.0	94308.80	628.725333
4	house	Yucatán	21.052583	-89.538639	205.0	105191.37	513.128634
5	house	Querétaro	20.716315	-100.452503	320.0	274034.68	856.358375
6	house	Morelos	18.812605	-98.954826	281.0	151509.56	539.179929
7	house	Chiapas	16.769737	-93.088928	140.0	79029.72	564.498000
8	house	Estado de México	19.305407	-99.646948	235.0	115937.75	493.352128
9	house	Morelos	18.804197	-98.932816	117.0	63223.78	540.374188

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

[11]:
```
VimeoVideo("656377991", h="c7319b0458", width=600)
```

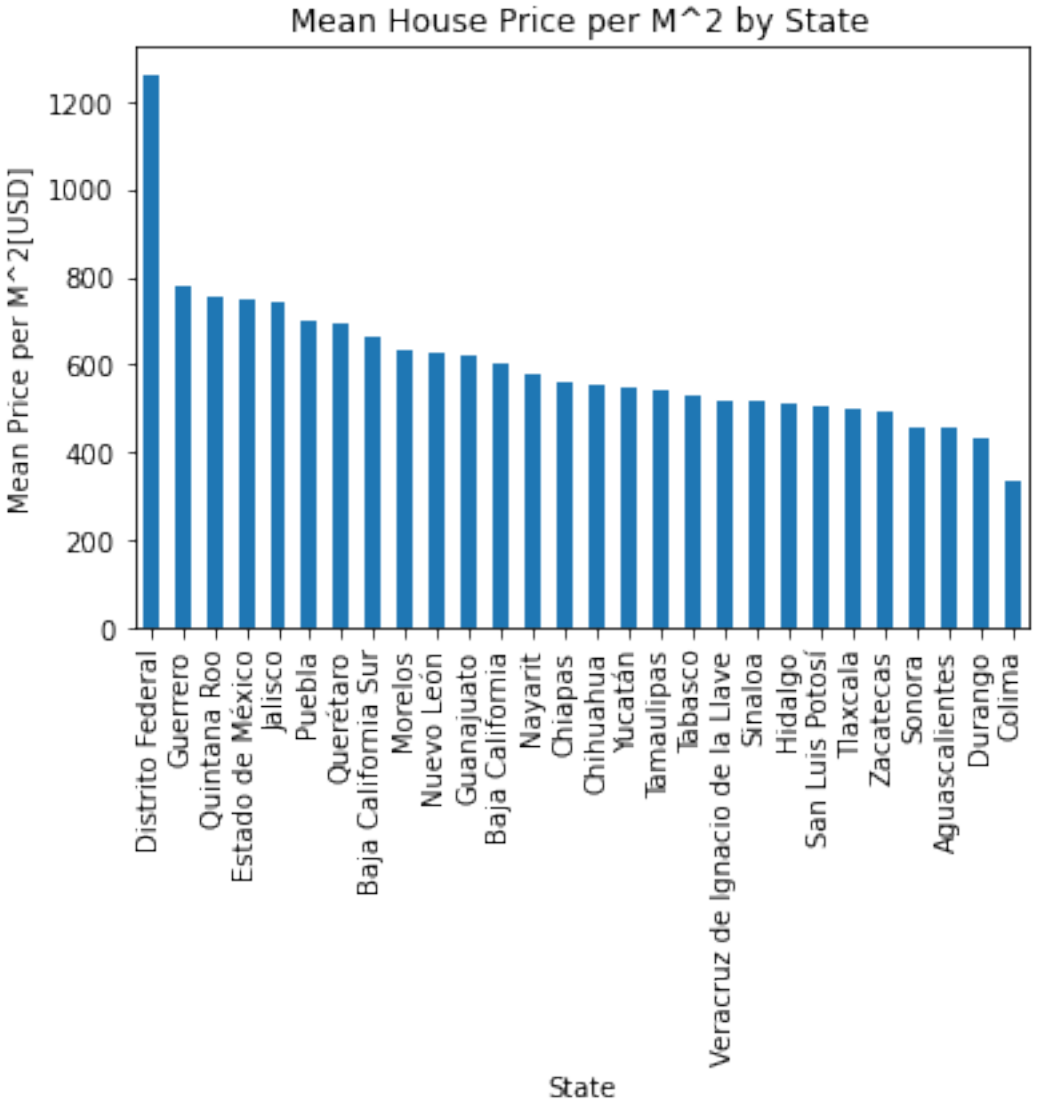
[11]:
```
<IPython.lib.display.VimeoVideo at 0x7f0f7b0b7ee0>
```

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.

[12]:
```
(
    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 Price per M^2 by State"
    )
);
```



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.

[13]:
```
VimeoVideo("656377758", h="62546c7b86", width=600)
```

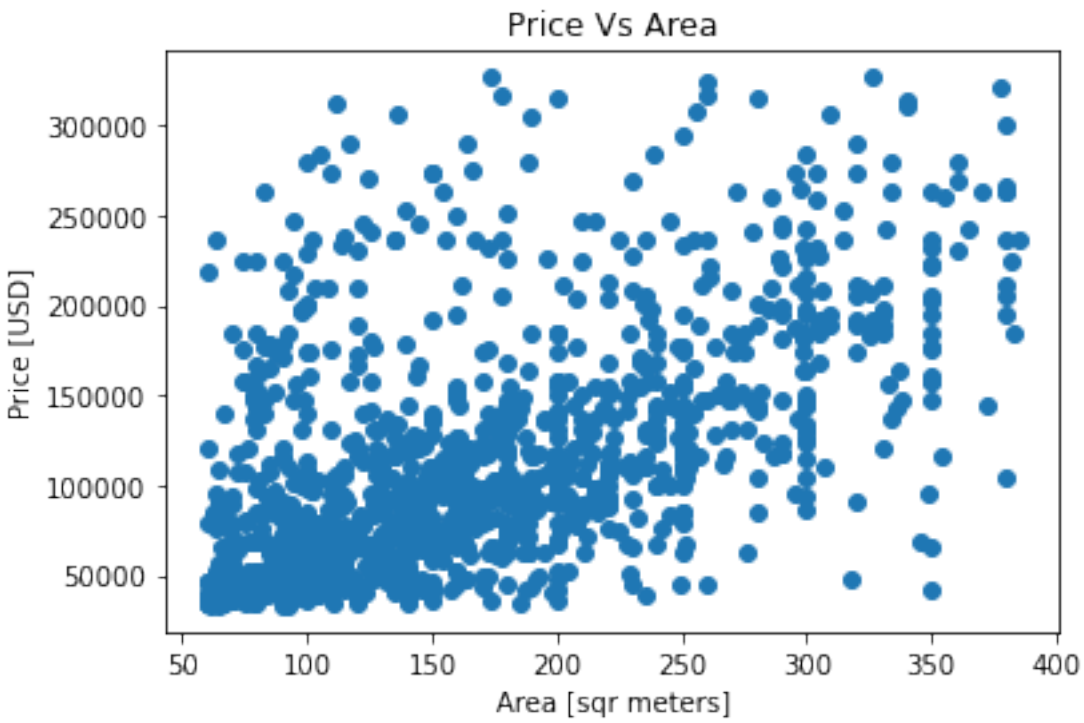
[13]:
```
<IPython.lib.display.VimeoVideo at 0x7f0f7af93df0>
```

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.

[14]:
```
plt.scatter(x=df["area_m2"], y=df["price_usd"])
plt.xlabel("Area [sqr meters]")
plt.ylabel("Price [USD]")
plt.title("Price Vs Area")
```

[14]:
```
Text(0.5, 1.0, '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?

[15]:
```
VimeoVideo("656377616", h="8d3b060e71", width=600)
```

[15]:
```
<IPython.lib.display.VimeoVideo at 0x7f0f7afaa1c0>
```

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.

[16]:
```
p_correlation = df["area_m2"].corr(df["price_usd"])
print(p_correlation)
```

```
0.5518728666998104
```

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.

[17]:
```
VimeoVideo("656377515", h="d2478d38df", width=600)
```

[17]:
```
<IPython.lib.display.VimeoVideo at 0x7f0f7dae79a0>
```

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.

[18]:
```
df_morelos = df[df["state"] == "Morelos"]
#df.shape
#df_morelos.shape
df_morelos.head()
```

[18]:
```
   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
```

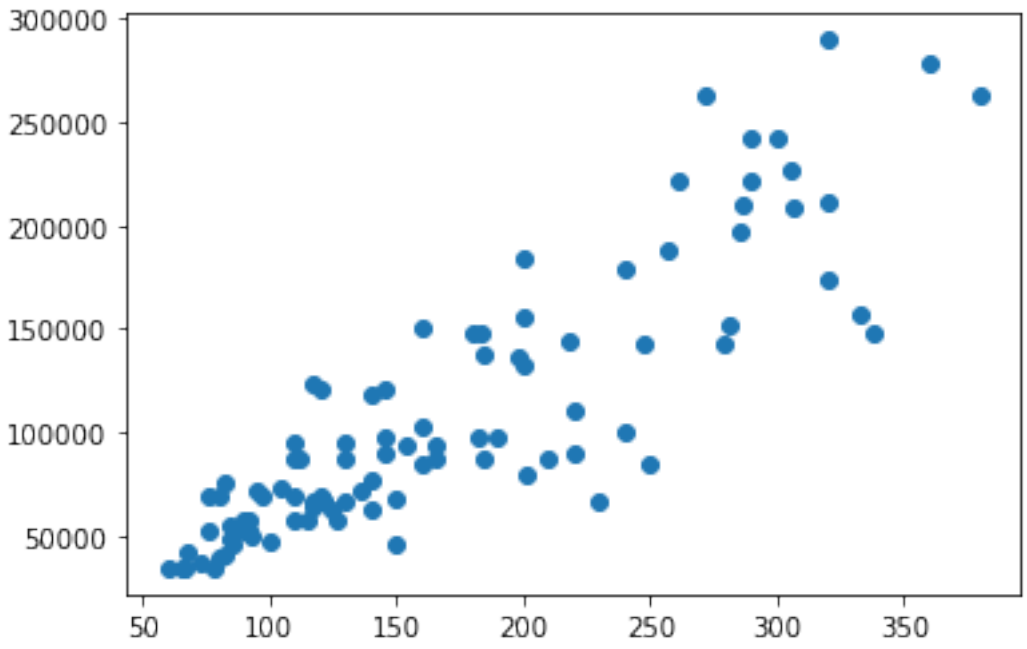
[19]:
```
VimeoVideo("656377395", h="bd93b05ff9", width=600)
```

[19]:
```
<IPython.lib.display.VimeoVideo at 0x7f0f7afaa8e0>
```

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.

[25]:
```
plt.scatter(x=df_morelos["area_m2"], y=df_morelos["price_usd"])
plt.xlabel ="Area [Square Metre]"
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.

[21]:
```
VimeoVideo("656377340", h="664cb44291", width=600)
```

[21]:
```
<IPython.lib.display.VimeoVideo at 0x7f0f7aef1190>
```

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.

[22]:
```
p_correlation = df_morelos["area_m2"].corr(df_morelos["price_usd"])
print(p_correlation)
```

```
0.8725659056131556
```

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*).

[23]:
```
VimeoVideo("656376911", h="19666a4c87", width=600)
```

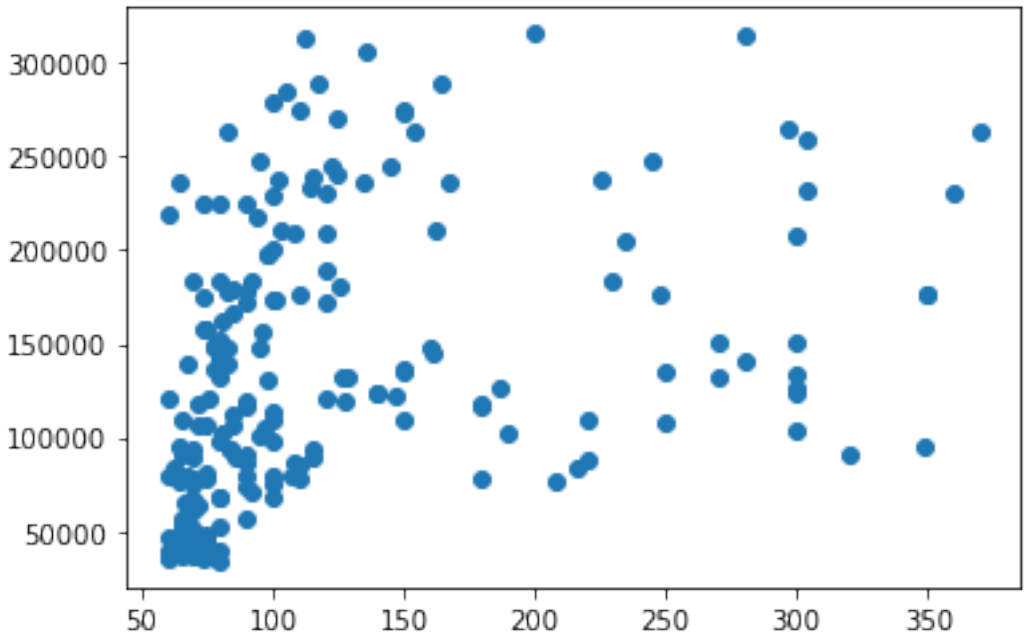
[23]:
```
<IPython.lib.display.VimeoVideo at 0x7f0f7aef1580>
```

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.

[28]:
```
# Subset `df` to include only observations from `"Distrito Federal"`
df_mexico_city = df[df["state"] == "Distrito Federal"]
#df_mexico_city.head();

# Create a scatter plot price vs area

plt.scatter(x=df_mexico_city["area_m2"], y=df_mexico_city["price_usd"])


plt.xlabel ="Area [Square Metre]"
plt.ylabel = "Price [USD]"
plt.title ="Mexico City: Price vs. Area"

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

```
0.34631963237093566
```



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.
