Location or Size: What Influences House Prices in Mexico?

In [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.

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.

In [2]:
```
df = pd.read_csv("data/Bikas-1st-concatenate-dataset.csv")
df.head()
```

Out[2]:

	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

In [3]:
```
df.describe()
```

Out[3]:

	lat	lon	area_m2	price_usd
count	1736.000000	1736.000000	1736.000000	1736.000000
mean	20.765410	-98.798575	170.261521	115331.980800
std	2.743425	4.882553	80.594539	65426.173793
min	15.752900	-117.054763	60.000000	33157.894737
25%	19.275200	-100.392553	101.750000	65789.473684
50%	19.620518	-99.204001	156.000000	99262.132105

	lat	lon	area_m2	price_usd
75%	21.073428	-98.245911	220.000000	150846.665000
max	32.665619	-86.767539	385.000000	326733.660000

In [ ]:
```
```

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.

In [4]:
```
VimeoVideo("656378731", h="8daa35d1e8", width=600)
```

Out[4]:

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.

In [5]:
```
mean_price_by_state = df.groupby("state")["price_usd"].mean().sort_values(ascending = F
mean_price_by_state
```

Out[5]:

```
state
Querétaro                          133955.913417
Guanajuato                         133277.965833
Nuevo León                         129221.985834
Distrito Federal                   128347.267365
Quintana Roo                       128065.415734
Chihuahua                          127073.851184
Jalisco                            123386.472237
Estado de México                   122723.490600
Campeche                           121734.633333
Puebla                             121732.974294
Guerrero                           119854.276015
Sonora                             114547.881798
Morelos                            112697.295615
Aguascalientes                     110543.888316
Baja California Sur                109069.339158
Yucatán                            108580.388526
Chiapas                            104342.313388
Veracruz de Ignacio de la Llave     96928.125254
Hidalgo                             94012.326563
Sinaloa                             93922.152490
Tamaulipas                          93713.386272
San Luis Potosí                     92435.540431
Nayarit                             87378.606842
Tabasco                             82763.586921
Durango                             78034.511729
Zacatecas                           76395.400000
Tlaxcala                            72921.819561
Colima                              65786.646947
Baja California                     63152.431198
Oaxaca                              59681.585000
Name: price_usd, dtype: float64
```

In [6]:

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

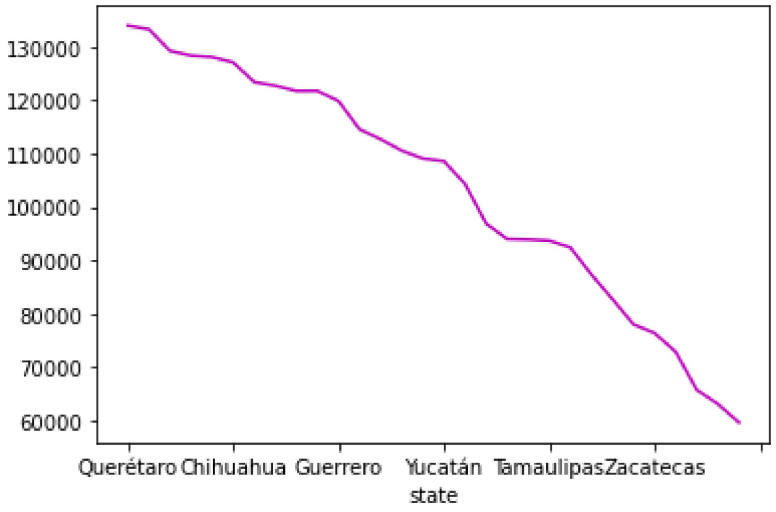
Out[6]:

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.

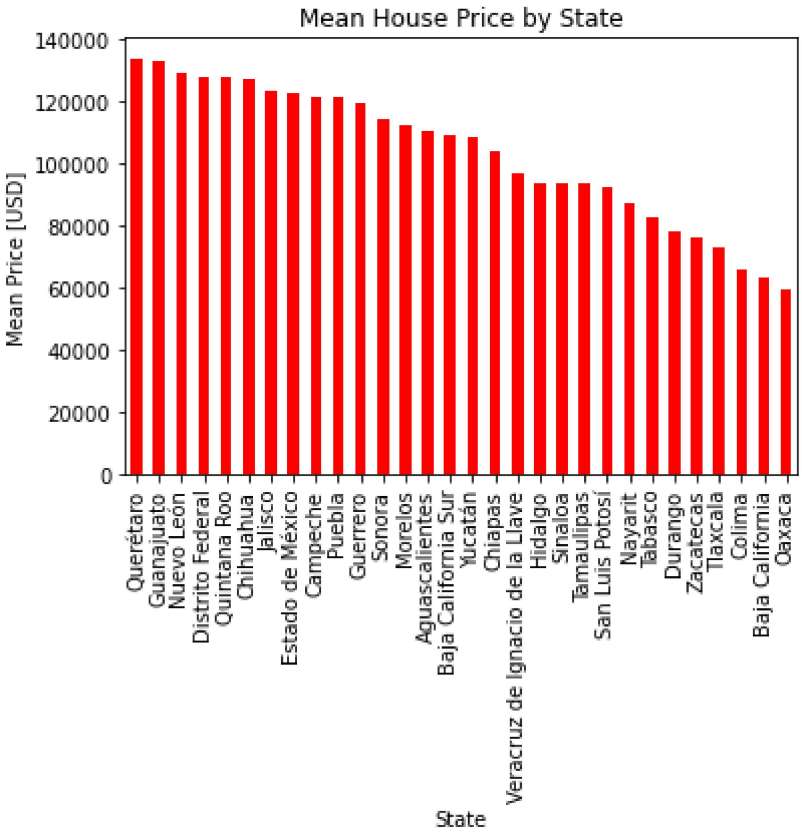
In [7]:
```
mean_price_by_state.plot(color = 'm')
```

Out[7]:
```
<AxesSubplot:xlabel='state'>
```



In [8]:
```
mean_price_by_state.plot(
    kind = "bar",
    xlabel = "State",
    ylabel = "Mean Price [USD]",
    title = "Mean House Price by State",
    color = 'r'

);
```



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 m^2. In order to do that, we need to create a new column.

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

Out[9]:

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.

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

Out[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

In [11]:
```
mean_price_by_state = df.groupby("state")["price_per_m2"].mean().sort_values(ascending
mean_price_by_state
```

Out[11]:
```
state
Distrito Federal                 1175.889150
Estado de México                  763.753423
Guerrero                          761.557207
Jalisco                           743.568106
```

```
Quintana Roo                        736.455283
Nuevo León                          723.710042
Puebla                              700.701977
Querétaro                           687.227849
Oaxaca                              683.019737
Guanajuato                          672.908100
Baja California Sur                 662.401955
Morelos                             649.383991
Tabasco                             643.503347
Campeche                            601.291762
Nayarit                             599.293638
Baja California                     598.852981
Chihuahua                           591.226745
Chiapas                             560.317147
Sinaloa                             546.536311
Yucatán                             545.889477
Tamaulipas                          541.282079
Veracruz de Ignacio de la Llave     531.536439
San Luis Potosí                     528.501600
Hidalgo                             527.970491
Zacatecas                           492.504078
Tlaxcala                            484.299623
Sonora                              461.845436
Aguascalientes                      449.859135
Durango                             415.447246
Colima                              355.110453
Name: price_per_m2, dtype: float64
```

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

In [12]:
```
VimeoVideo("656377991", h="c7319b0458", width=600)
```

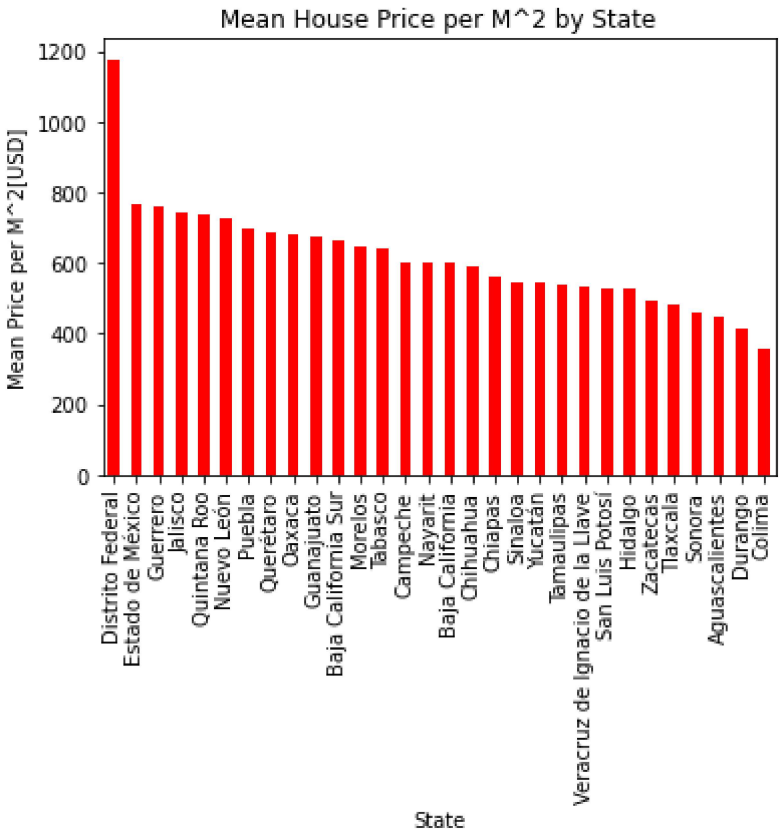
Out[12]:

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 m^2 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.

In [13]:
```
(
    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",
        color = 'r'
    )
);
```



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.

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.

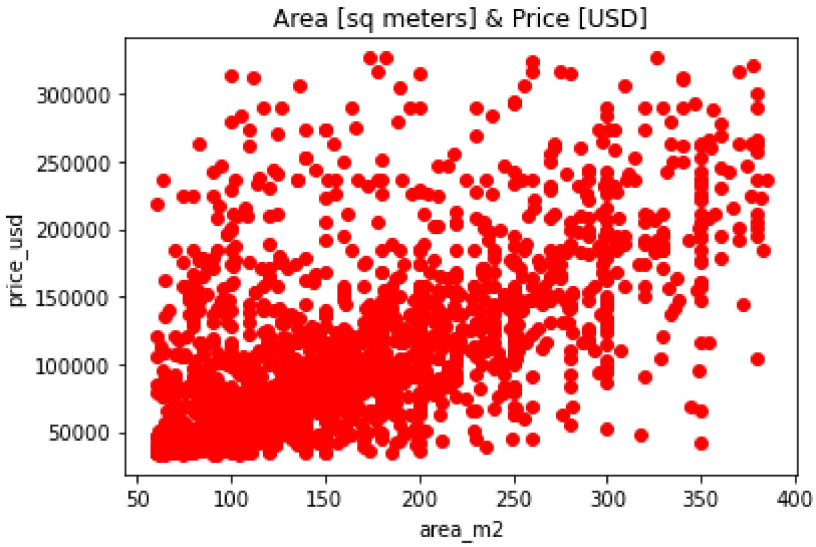
In [14]:
```
VimeoVideo("656377758", h="62546c7b86", width=600)
```

Out[14]:

price_usd**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.

In [28]:
```
plt.scatter(df["area_m2"], df["price_usd"], color = "r")
plt.xlabel("area_m2")
plt.ylabel("price_usd")
plt.title("Area [sq meters] & Price [USD]");
```



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?

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

Out[15]:

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.

In [29]:
```
p_correlation = df["area_m2"].corr(df["price_usd"])
print(p_correlation)
```

```
0.5855182454266904
```

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.

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

Out[17]:

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.

In [31]:
```
df_morelos = df[df["state"] == "Morelos"]
df_morelos.shape
```

Out[31]:
```
(160, 7)
```

In [32]:
```
df_morelos.head()
```

Out[32]:

	property_type	state	lat	lon	area_m2	price_usd	price_per_m2
6	house	Morelos	18.812605	-98.954826	281.0	151509.56	539.179929
9	house	Morelos	18.804197	-98.932816	117.0	63223.78	540.374188
18	house	Morelos	18.855343	-99.241142	73.0	36775.16	503.769315
49	house	Morelos	18.804197	-98.932816	130.0	65858.10	506.600769
55	house	Morelos	18.960244	-99.212962	305.0	227351.46	745.414623

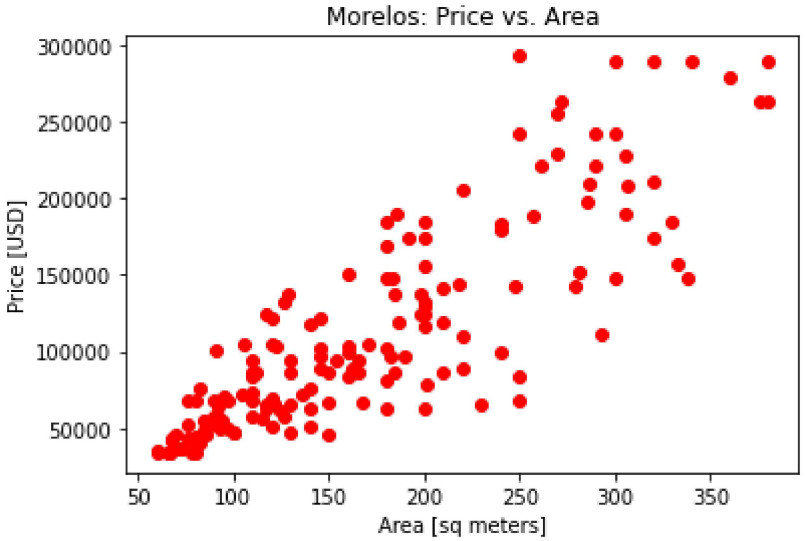
In [19]:
```
VimeoVideo("656377395", h="bd93b05ff9", width=600)
```

Out[19]:

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.

In [35]:
```
plt.scatter(df_morelos["area_m2"], df_morelos["price_usd"], color = "r")
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.

In [20]:
```
VimeoVideo("656377340", h="664cb44291", width=600)
```

Out[20]:

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.

In [36]:
```
p_correlation = df_morelos["area_m2"].corr(df_morelos["price_usd"])
print(p_correlation)
```

```
0.8498077614061482
```

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*).

In [22]:
```
VimeoVideo("656376911", h="19666a4c87", width=600)
```

Out[22]:

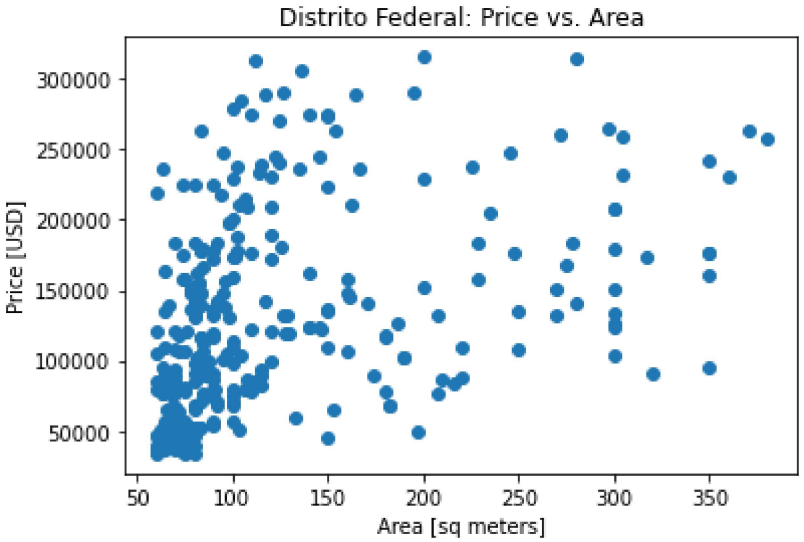
Task 1.4.11: First, create a new DataFrame called `df_mexico_city` that includes all the observations from `df` that are part of the *Distrito Federal*. Next, create a scatter plot that shows price vs area. Don't forget to label the x- and y-axis and use the title `"Mexico City: Price vs. Area"`. Finally,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

In [41]:
```
# 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(df_mexico_city["area_m2"], df_mexico_city["price_usd"], color )
plt.xlabel("Area [sq meters]")
plt.ylabel("Price [USD]")
plt.title("Distrito Federal: Price vs. Area");
p_correlation = df_mexico_city["area_m2"].corr(df_mexico_city["price_usd"]).round(4)
print(p_correlation)
```

```
0.41
```



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.
