Markdown

Python 3 (ipykernel)

"lon", respectively.

- Drop rows with missing values from a DataFrame using pandas.
- Split the strings in one column to create another using pandas.

[11]:
```
df3.dropna(inplace=True)
df3[["lat", "lon"]]= df3["lat-lon"].str.split(",", expand=True)
df3.head()
```

[11]:

	property_type	place_with_parent_names	lat-lon	area_m2	price_usd	lat	lon
0	apartment	|México|Distrito Federal|Gustavo A. Madero|Acu...	19.52589,-99.151703	71.0	48550.59	19.52589	-99.151703
1	house	|México|Estado de México|Toluca|Metepec|	19.2640539, -99.5727534	233.0	168636.73	19.2640539	-99.5727534
2	house	|México|Estado de México|Toluca|Toluca de Lerd...	19.268629,-99.671722	300.0	86932.69	19.268629	-99.671722
4	apartment	|México|Veracruz de Ignacio de la Llave|Veracruz|	19.511938,-96.871956	84.0	68508.67	19.511938	-96.871956
5	house	|México|Jalisco|Guadalajara|	20.689157,-103.366728	175.0	102763.00	20.689157	-103.366728

[12]:
```
VimeoVideo("656314050", h="13f6a677fd", width=600)
```

Markdown

Python 3 (ipykernel)

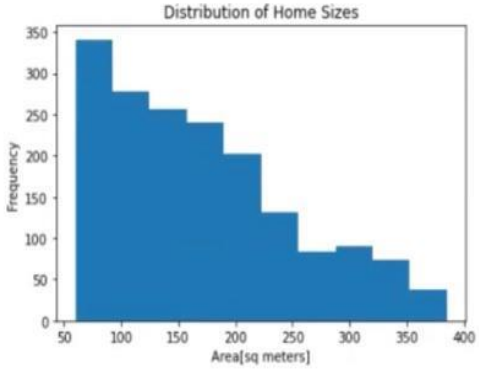
[11]:

	area_m2	price_usd
count	1736.000000	1736.000000
mean	170.261521	115331.980766
std	80.594539	65426.173873
min	60.000000	33157.890000
25%	101.750000	65789.470000
50%	156.000000	99262.130000
75%	220.000000	150846.665000
max	385.000000	326733.660000

Markdown

Python 3 (ipykernel)

[19]:
```
plt.hist(df["area_m2"])
plt.xlabel("Area[sq meters]")
plt.ylabel("Frequency")
plt.title("Distribution of Home Sizes");
```



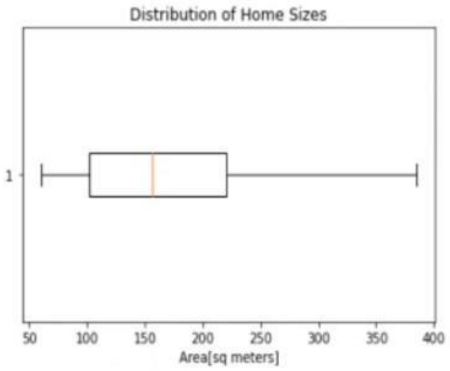
Looking at our histogram, we can see that `"area_m2"` skews right. In other words, there are more houses at the lower end of the distribution (50–200m^2) than at the higher end (250–400m^2). That explains the difference between the mean and the median.

Markdown

Python 3 (ipykernel)

- Create a boxplot using Matplotlib.

[24]:
```
plt.boxplot(df["area_m2"], vert=False)
plt.xlabel("Area[sq meters]")
plt.title("Distribution of Home Sizes");
```



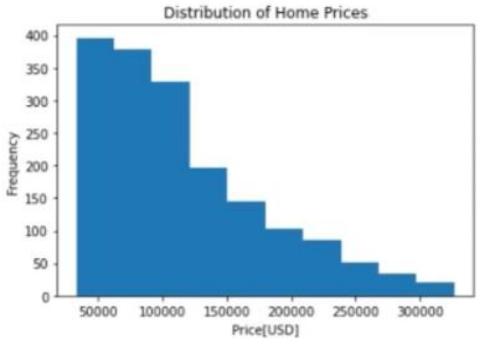
Does `"price_usd"` have the same distribution as `"price_per_m2"` ? Let's use the same two visualization tools to find out.

Markdown

Python 3 (ipykernel)

- Create a histogram using Matplotlib.

[27]:
```
plt.hist(df["price_usd"])
plt.xlabel("Price[USD]")
plt.ylabel("Frequency")
plt.title("Distribution of Home Prices");
```



Looks like `"price_usd"` is even more skewed than `"area_m2"` . What does this bigger skew look like in a boxplot?

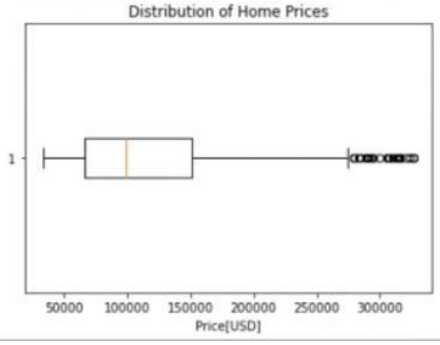
Markdown

Python 3 (ipykernel)

- What's an outlier?
- Create a boxplot using Matplotlib.

[30]:
```
plt.boxplot(df["price_usd"], vert=False)
plt.xlabel("Price[USD]")
plt.title("Distribution of Home Prices")
```

[30]:
```
Text(0.5, 1.0, 'Distribution of Home Prices')
```



Markdown

Python 3 (ipykernel)

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

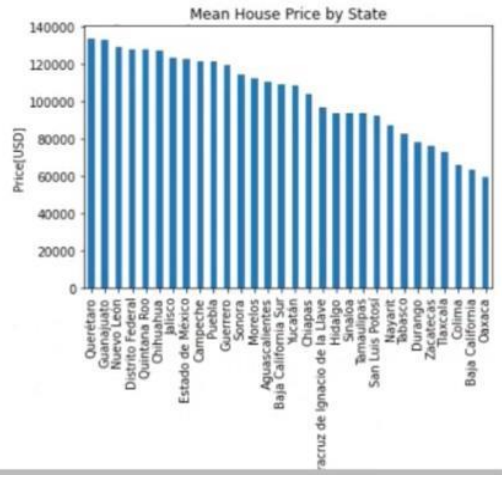
[4]:
```
mean_price_by_state = df.groupby("state") ["price_usd"].mean().sort_values(ascending=False)
mean_price_by_state
```

[4]:
```
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
```

Markdown

Python 3 (ipykernel)

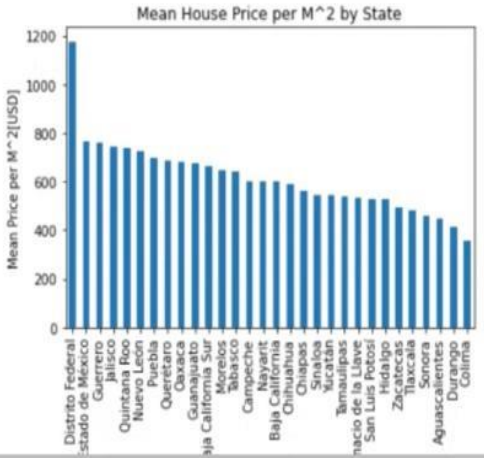
```
    title="Mean House Price by State"
);
```



Markdown

Python 3 (ipykernel)

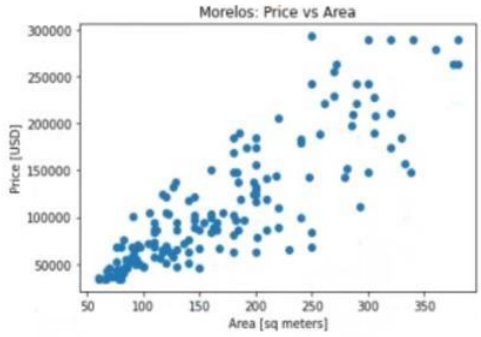
```
        xlabel="States",
        ylabel="Mean Price per M^2[USD]",
        title="Mean House Price per M^2 by State"
    )
);
```



Markdown

Python 3 (ipykernel)

[26]:
```
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");
```



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"`.

Markdown

Python 3 (ipykernel)

```
plt.ylabel("Price [USD]")
plt.title("Price vs Area");
```

Markdown

Python 3 (ipykernel)

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

```
0.41070392130717887
```