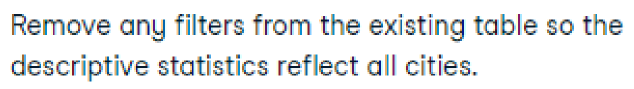
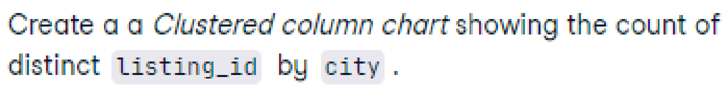
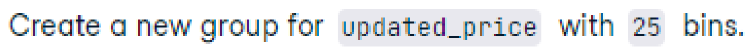
Distributions and Outliers Lab

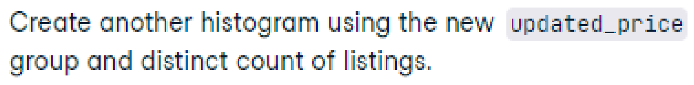
Load airbnb dataset: https://github.com/fenago/cts245X/tree/main/EDA

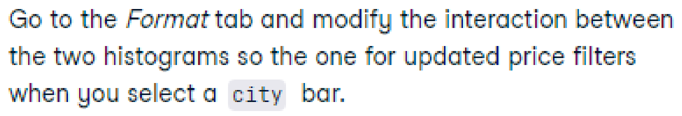
Part 1: Create your first histogram

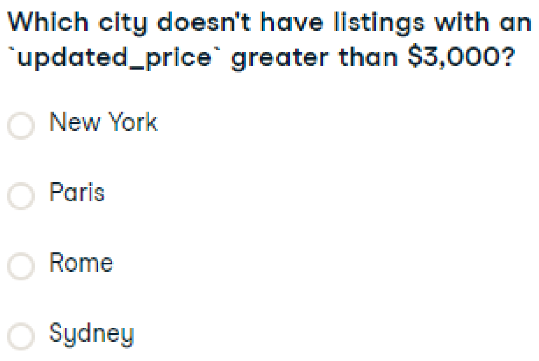

Remove any filters from the existing table so the descriptive statistics reflect all cities.


Create a a *Clustered column chart* showing the count of distinct `listing_id` by `city` .


Create a new group for `updated_price` with `25` bins.


Create another histogram using the new `updated_price` group and distinct count of listings.


Go to the *Format* tab and modify the interaction between the two histograms so the one for updated price filters when you select a `city` bar.


Which city doesn't have listings with an `updated_price` greater than $3,000?

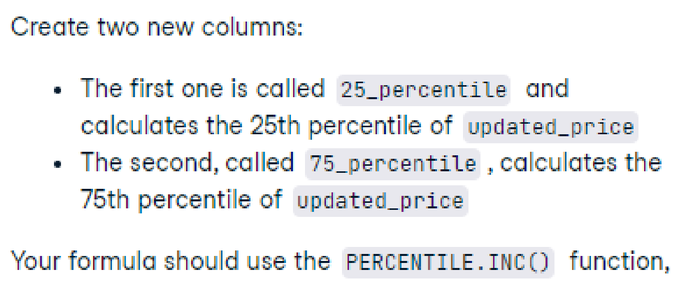
○ New York

○ Paris

○ Rome

○ Sydney

Part 2: Identify outliers

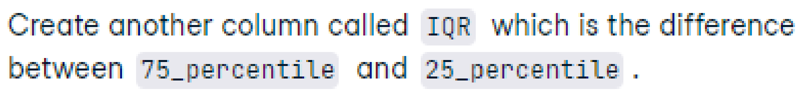

Create two new columns:

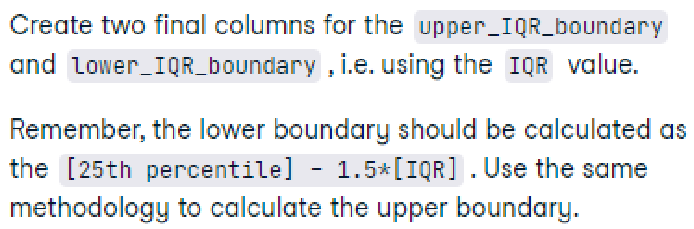
- The first one is called `25_percentile` and calculates the 25th percentile of `updated_price`
- The second, called `75_percentile`, calculates the 75th percentile of `updated_price`

Your formula should use the `PERCENTILE.INC()` function,

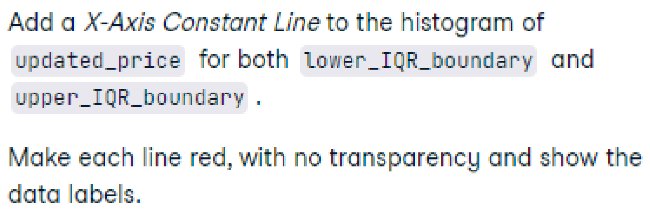
For more information visit this documentation.

https://learn.microsoft.com/en-us/dax/percentile-inc-function-dax

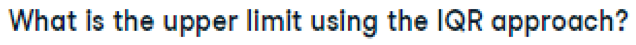

Create another column called `IQR` which is the difference between `75_percentile` and `25_percentile`.


Create two final columns for the `upper_IQR_boundary` and `lower_IQR_boundary`, i.e. using the `IQR` value.

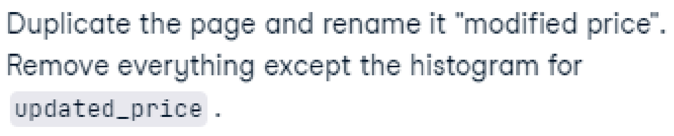
Remember, the lower boundary should be calculated as the `[25th percentile] - 1.5*[IQR]`. Use the same methodology to calculate the upper boundary.


Add a *X-Axis Constant Line* to the histogram of `updated_price` for both `lower_IQR_boundary` and `upper_IQR_boundary`.

Make each line red, with no transparency and show the data labels.


What is the upper limit using the IQR approach?

Part 3: Addressing outliers in the data


Duplicate the page and rename it "modified price". Remove everything except the histogram for `updated_price`.

Create a column called `modified_price` using nested `IF()` statements. The final structure should look like:

```
modified_price = IF( ___ < PERCENTILE.INC(___,
0.05), PERCENTILE.INC(___, 0.05), IF( ___ >
PERCENTILE.INC(___, 0.95), PERCENTILE.INC(___,
0.95), ___ ))
```

Add a new table with a distinct count of listings and `modified_price` as a median and average.

Create a new group with 25 bins of `modified_price` .

Create a histogram for `modified_price` .

How many distinct listings are in the last bin?