DATA SCIENCE PROJECT

OVERALL TASK: **In this project, you'll work with a dataset of homes for sale in Brazil. Your goal is todetermine if there are regional differences in the real estate market. Also, you will look at southernBrazil to see if there is a relationship between home size and price.**

Task 1: DATA CLEANING of `brasil-real-estate-1.csv` raw data.

Task 1.1: Import the CSV file into the DataFrame `df1`.

Task 1.2: Before you move to the next task, take a moment to inspect df1 using the info and head methods. What issues do you see in the data? What cleaning will you need to do before you can conduct your analysis?

Task 1.3: Use the `"lat-lon"` column to create two separate columns in `df1`: `"lat"` and `"lon"`. Make sure that the data type for these new columns is `float`.

Task 1.4: Use the `"place_with_parent_names"` column to create a `"state"` column for `df1`. (Note that the state name always appears after `"|Brasil|"` in each string.)

Task 1.5: Transform the `"price_usd"` column of `df1` so that all values are floating-point numbers instead of strings.

Task 1.6: Drop the `"lat-lon"` and `"place_with_parent_names"` columns from `df1`.

Task 1.7: Give a short report on the data you now have.

TASK 2: DATA CLEANING of `brasil-real-estate-2.csv` raw data.

Task 2.1: Import the CSV file `brasil-real-estate-2.csv` into the DataFrame `df2`

Task 2.2: Before you jump to the next task, look at `df2` using the `info` and `head` methods. What issues do you see in the data? How is it similar or different from `df1`?

Task 2.3: Use the `"price_brl"` column to create a new column named `"price_usd"`. (Keep in mind that, when this data was collected in 2015 and 2016, a US dollar cost 3.19 Brazilian reals.)

Task 2.4: Drop the `"price_brl"` column from `df2`, as well as any rows that have `NaN` values.

Task 2.5: Concatenate `df1` and `df2` to create a new DataFrame named `df`.

TASK 3: EXPLORATION - In this section, you'll use your new data visualization skills to learn more about the regional differences in the Brazilian real estate market.

TASK 3.1: Use the `describe` method to create a DataFrame `summary_stats` with the summary statistics for the `"area_m2"` and `"price_usd"` columns.

TASK 3.2: Create a histogram of `"price_usd"`. Make sure that the x-axis has the label `"Price [USD]"`, the y-axis has the label `"Frequency"`, and the plot has the title `"Distribution of Home Prices"`.
TASK 3.3: Create a horizontal boxplot of `"area_m2"`. Make sure that the x-axis has the label `"Area [sq meters]"` and the plot has the title `"Distribution of Home Sizes"`.

TASK 3.4: Use the `groupby` method to create a Series named `mean_price_by_region` that shows the mean home price in each region in Brazil, sorted from smallest to largest.

TASK 3.5: Use `mean_price_by_region` to create a bar chart. Make sure you label the x-axis as `"Region"` and the y-axis as `"Mean Price [USD]"` and give the chart the title `"Mean Home Price by Region"`.

You're now going to shift your focus to the southern region of Brazil and look at the relationship between home size and price.

TASK 3.7: Create a DataFrame `df_south` that contains all the homes from `df` that are in the `"South"` region.

TASK 3.8: Use the `value_counts` method to create a Series `homes_by_state` that contains the number of properties in each state in `df_south`.

TASK 3.9: Create a scatter plot showing price vs. area for the state in `df_south` that has the largest number of properties. Be sure to label the x-axis `"Area [sq meters]"` and the y-axis `"Price [USD]"`; and use the title `"<name of state>: Price vs. Area"`

TASK 3.9.1: Create a dictionary `south_states_corr`, where the keys are the names of the three states in the `"South"` region of Brazil, and their associated values are the correlation coefficient between `"area_m2"` and `"price_usd"` in that state.

As an example, here's a dictionary with the states and correlation coefficients for the Southeast region. Since you're looking at a different region, the states and coefficients will be different, but the structure of the dictionary will be the same.

```
{'Espírito Santo': 0.6311332554173303,

 'Minas Gerais': 0.5830029036378931,

 'Rio de Janeiro': 0.4554077103515366,

 'São Paulo': 0.45882050624839366}
```

SUBMISSION REQUIREMENTS:

1. 1 zipped folder including:

a. An executable Jupyter notebook that includes the code, its output, and the answer to each question along with the solution.

b. A duplicate version of the Jupyter notebook above in PDF to include the output of the code, you must RUN the code before downloading the PDF.