The first step was importing the listing data from the local csv file as a pandas dataframe named “listing\_data”. I then performed the following steps inspect the data after importing the data

1. Inspect the data by calling the .head() and .tail() method to inspect the dataframe which showed the first and last 5 rows and all the columns of the dataframe. However, there are 95 columns in the dataframe and therefore some of the middle columns in the dataframe are not displayed
2. Call the .columns attribute to inspect and get a list of the 95 columns in the dataframe
3. Call the .shape attribute to inspect the number of rows and columns in the dataframe get to know that there are 8706 rows and 95 columns in the dataframe
4. Call the .info() method to find out which columns have missing data points and suspicious datatypes
   1. There are 3 empty columns with no data: host\_acceptance\_rate, neighbourhood\_group\_cleansed, has\_availability
   2. The datatype of the “price” columns is string instead of float
   3. There are numerous columns missing data. However, I believe they are fields that do not necessarily require a value such a summary of the listing, description of access and transit which is up to the discretion of the host to put in values
5. Convert the datatype of the “price” column from string to float first removing the leading “$” and the “,” within the string and then call pd.to\_numeric to convert to float64 datatype
6. Call the .describe() method on the “listing\_data” dataframe to compute the summary statistics for all the numeric columns. Summary statistics for 33 numeric columns are returned. However, some of the columns are not displayed due to not enough room.
7. Inspecting the summary statistics and plotting the “price” column tells me that the distribution of the price per nice is heavily skewed to the right. The median price per night is at $160 but the maximum is at $10,000 while the minimum is at $0.
8. There is one listing with price $0 and it looks like it is a test listing based on the summary so I decided to drop it from “listing\_data” dataframe.
9. Call the .describe() method to inspect the new minimum price per night and it is $10. This still looks low so I decided to take a look at all the listings that are under $30 per night and it seems that a lot of them are shared room listing.
10. Taking a closer look at the listing by room type shows that there are only less than 200 listing that are shared room listings out of over 8000 listing. I decided to focus my analysis on listings that are for entire home or private room instead of shared room and therefore removed the shared room listings from the “listing\_data” data
11. After removing the shared room listing from the data, there were still 5 listing with $10 per night. 4 of them were specified as test listings in either the name of description of the listing. The price for the last one seems too low to be true either so I decided to remove it as well.
12. Taking a look at the extreme pricey listings (i.e. over $9,000 per night) reveals a similar pattern. There are 9 such listings and 6 of them a obviously invalid based on just the name and description of the listings
13. A lot of the listings with extremely high and low price that I just looked at seemed to be listing set for testing purpose. This made me wonder if there are other test listing out there that were set up for test purpose. I tried to locate such listing for by filtering the listings containing “test” or “Test in their name and found 11 of such listings
14. A lot of the test or unreal listings have no reviews and no availability in the next 365 days. I am wondering if I should remove all such listings and only analyze listings that have at list one comment or some availabilities in the next 365 days. However, there are 1153 listings that have no reviews and no availability in the next 365 days. I am not sure if I should remove such a large chunk of data.