Group Members: Devin Taing, Michael huh, Steve Nguyen, Ji Soo Kim

1. **Identify missing values**
    - **Note that there are trailing spaces/new lines. How to handle it?**
        i. Remove all trailing spaces, new lines, and any white space to make the data more uniform and easier to work with. For example, you don't have to worry about different amounts of spaces when searching for strings or other operations like that.
    - Since the dataset is a .csv (comma separated values) file, any missing values are going to be denoted with two consecutive commas. This is because each data point is separated by a comma, so two commas in a row means that there is a value missing from between them.
2. **Remove rows with missing values**
    - **Or is it better to replace the missing values with NA or another value? Try writing scripts for both methods depending on the field of interest.**
    - awk -F, '{for(i=1;i<=NF;i++)if($i==""){next}}1' test.csv > full.csv
        i. skips rows with missing values
            1. We decided to do this because we have a very large dataset and removing a small percentage of data wouldn't affect the final result.
3. **Remove duplicate entries**
    - Duplicate entries can skew a dataset's averages or medians with data points that were already considered, so there is no reason to have duplicate entries.
        i. sort -u -n -t, -k 10,10 full.csv > unique.csv
4. **Identifying and handling outliers**
    - **How to identify outliers? Calculate the mean/median/or mode and use that as the condition**
    - IQR rule/method:
        i. Calculate Q1, Q3, and the IQR
            1. Q1: (n + 1) x 0.25
            2. Q3: (n + 1) x 0.75
            3. IQR: Q3 - Q1
        ii. Multiply the IQR by 1.5 and subtract this from Q1 and add this to Q3
        iii. Any datapoint that falls outside of this range (Q1 - 1.5IQR to Q3 + 1.5IQR) is considered an outlier
        iv. The IQR focuses on the middle 50% of values, so any outliers found following this method will be far outside from the rest of the data
5. **After you validate each set, write a shell script to automate. We are assuming we will merge the NY dataset with other datasets from AirBnb from other regions.**