Data Wrangling

Collecting Data:
The data was sourced from http://insideairbnb.com/get-the-data.html which is an independent, non-commercial website that utilizes publicly available information about San Francisco's Airbnbs. The listings.csv.gz file was obtained and the raw dataset contained 17 columns and 8533 records. The data was compiled on December 4th, 2019.

Selected Columns:

- id
- host_id
- zipcode
- property_type
- room_type
- accommodates
- bathrooms
- bedrooms
- beds
- bed_type
- amenities
- price
- minimum_nights
- availability_30
- number_of_reviews
- review_scores_rating
- calculated_host_listings_count

Removed Columns:

- listing_url
- scrape_id
- last_scraped
- name
- summary
- description
- experiences_offered
- neighborhood_overview
- notes
- transit

- access
- interaction
- house_rules
- thumbnail_url
- medium _url
- picture_url
- xl_picture_url
- host_url
- host_name
- host_since
- host_location
- host_about
- host_response _time
- hos_response_rate
- host_acceptance_rate
- host_is_superhost
- host_thumbnail_url
- host_picture_url
- host_neighbourhood
- host_listings_count
- host_total_listings_count
- host_verifications
- host_has_profile_pic
- host_identity _verified
- street
- neighbourhood
- neighbourhood_cleansed
- neighbourhood_group_cleansed
- city
- state
- market
- smart_location
- country_code
- country
- latitude
- longitude
- is_location_exact
- square_feet
- Weekly_price

- monthly_price
- security_deposit
- cleaning_fee
- guests_included
- extra_people
- minimum_nights
- minimum_maximum_nights
- maximum_minimum_nights
- minimum_maximum_nights
- maximum_maximum_nights
- minimum_nights_avg_ntm
- maximum_nights_avg_ntm
- calendar_updated
- has_availability
- availability_60
- availability_90
- availability_365
- calendar_last_scraped
- number_of_reviews_ltm
- first_review
- last_review
- review_scores_accuracy
- review_scores_cleanliness
- review_scores_checkin
- review_scores_communication
- review_scores_location
- review_scores_value
- requires_license
- license
- jurisdiction _names
- instant_bookable
- is_business_travel_ready
- cancellation_policy
- require_guest_profile_picture
- require_guest_phone_verification
- calculated_lost_listings_count_entire_homes
- calculated_lost_listings_count_private_homes
- calculated_lost_listings_count_shared_homes
- Reviews_per_month

A Pandas Dataframe was created with the selected columns and null values were found in 5 columns:

- review_scores_rating: 1,932 records, or 22.6%
- zipcode: 245 records, or .03%
- beds: 12 records, or .00%
- bathrooms: 5 records, or .00%
- bedrooms: 3 records, or .00%

The total amount of the null values are negligible, so these records were dropped from the dataframe.

Price:

The price column needed to be cleaned up and transformed into a float data type. The '$' and ',' were removed from the values, thus enabling the column to be changed to float type.

Zipcode:

The zipcode column contained a few data irregularities. "CA " was removed from the beginning of the values and the records listed as "CA" were dropped as well. **How many CA values were removed?? The column was then transformed into integer type.

Lastly, all numerical columns (accommodates, bathrooms, bedrooms, beds, price) with values of zero were removed from the dataframe. The data is now clean and ready for analysis!