This set contains Airbnb listing information in US cities, though information about most major cities around the world is available. This is more immediately interesting to me, as I’m from the US. I would more than likely dive into the worldwide information on my own later, but I have plenty of free time data sets to play as a result of the hunting I did for this assignment.

DATA SOURCE

<https://www.kaggle.com/kritikseth/us-airbnb-open-data>

This data set was organized by user [Kritik Seth](https://www.kaggle.com/kritikseth) on Kaggle and sourced from [Inside Airbnb](http://insideairbnb.com/).

**Inside Airbnb** is an independent organization that seeks to disseminate information about [Airbnb](https://www.airbnb.com/) and its effect on local residential markets. They are not directly affiliated with Airbnb, however the information they collect is scraped directly from Airbnb. This is an open data set.

SUMMARY

The initial data set was created via web scraping of the Airbnb website. There are some possible issues here, a faulty line in the scraper program could result in inaccurate information. However, after going over it I feel like this is not the case in the instance. There are some issues with observations in the name column, however in the interest of user privacy I’ll be dropping that column. The data itself covers the period of Oct 2019-Oct 2020 as Inside Airbnb only retains 12 months of data at any given time. Which brings me to my only real concern as far as bias goes, which is that they year 2020 was not exactly one filled with travel. As such I would expect this particular data set to be a little lighter than other years. This is something to be explored later however, as in order to get other dates I would need to run a formal request with Inside Airbnb and that could take longer (and be more expensive) than I have time for.

Limitations

The limitations to this data set are that it is scraped from Airbnb and only goes back 12 months. These are limitations because while the information is cleaned before it placed on the Inside Airbnb website, it clearly isn’t perfect. There are limitations to what a web scraper can do, as it can’t really interpret what is or isn’t important information to collect, it can only do exactly what it is told to do. So, anything shortcomings in the code written to create the scraper would be reflected in the data retrieved. That said it is fairly comprehensive. Inside Airbnb claims to collect 50 columns of data points, there are only 16 here. So, some fields have been left out. The other concern is the timeliness of the data. It is from October 2019 to October 2020. Which is cool, but it isn’t actually listed anywhere. While that isn’t a problem per say, I would argue that *with* that information I could see how long listings have been posted before they were taken down or when they were put up. A dimension of data that could prove interesting. Still, what we have is plenty and good to work with. I find I can get a little greedy when it comes to how much information I want to sift through.

PROFILE

# Cleaning

Luckily, I didn’t have a *ton* of things to do here So I’ll just go over them quickly.

### Data Type Consistency Check

There was an issue with the data types being inconsistent in a few columns. Specifically, name, host\_name, neighborhood\_group, and last\_review. Along side that the columns id, and host\_id were both listed as integers. So, I changed all of those into strings, as there was no mathematical value to the numbers in any of those places. I didn’t end up changing latitude or longitude because they were listed as floats, and I didn’t really want to accidently change that somehow as I’ll need it later one to be sure. I’m sure I could have changed the type to a string, I was just being overly cautious.

### Missing Values Check

There was an issue with missing values in five columns: name, host\_name, neighbourhood\_group, last\_review, and reviews\_per\_month. The missing values for last\_review and reviews\_per\_month were the same, so I looked into number\_of\_reviews and found that they correlated with number\_of\_reviews being 0. So, I changed the values to “0” for reviews per month and “None” in last\_review. Where the value in name was missing, I ultimately removed them. Though this proved a touch more challenging that I was expecting. For some reason when I did the dropna() function it removed *all* the values from the column. I just reloaded the kernel and did it the long way but it was annoying. As for host\_name, I was going to drop the column, so I wasn’t worried about that.

### Duplicates Check

I found one duplicate and got rid of it. I found it by checking the value counts for listing\_id. There was one with 2 occurrences and there should have been *none* since it’s a unique number per listing. Not sure if it was a collection error or technical one but I removed it either way.

### Final Adjustments

Lastly, I went through and just changed id and name to listing\_id and listing\_name. Then moved city over so it wasn’t just at the end where it made no sense to be.

# Columns

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column Name | Description | Data Type | Min Value | Max Value | Average Value |
| listing\_id | unique id for listing | str | n/a | n/a | n/a |
| listing\_name | user entered name for listing | str | n/a | n/a | n/a |
| host\_id | unique id for host | str | n/a | n/a | n/a |
| neighbourhood\_group | name of grouped neighborhoods | str | n/a | n/a | n/a |
| neighbourhood | name of neighborhood | str | n/a | n/a | n/a |
| latitude | co-ordinates | float64 | n/a | n/a | n/a |
| longitude | co-ordinates | float64 | n/a | n/a | n/a |
| city | name of city | str | n/a | n/a | n/a |
| room\_type | room type code | str | n/a | n/a | n/a |
| price | listed price per night | int64 | 0.00 | 24999.00 | 219.72 |
| minimum\_nights | fewest nights to stay | int64 | 1 | 1250 | 10.12 |
| number\_of\_reviews | total number of reviews | int64 | 0 | 966 | 34.51 |
| last\_review | date of most recent review | str | n/a | n/a | n/a |
| reviews\_per\_month | total reviews divided by 12 | float64 | 0 | 44.06 | 1.12 |
| calculated\_host\_listings\_count | total number of listings per host | int64 | 1 | 593 | 16.7 |
| availability\_365 | number of listings available days | int64 | 0 | 365 | 159.33 |

QUESTIONS

* Is there a correlation between accounts with a significant number of listings and the prices they charge?
  + Check neighborhoods where there is a mix of account listings and compare pricing
* Which regions are more expensive to rent?
  + East coast, west coast, central
* What city has the most listings?
  + Top/Bottom 5
* Room types probably have a significant impact on pricing
  + Room type distribution? What type is concentrated where?
* Minimum number of nights to stay gets WAY up there
  + Is there a consistent location?
  + Or maybe certain users?
    - Users with lots of listings, maybe putting some on a soft hold?
    - Same with that 25k a night one?
* Does the number of reviews a listing have seem to have any relation with the cost?
  + This would be easier to see if we had a time series on the number of reviews a give listing has but failing that we could just see if listings with more reviews tend to be more or less expensive than the average
* Does availability reflect anything in particular?
  + Could there be a connection between days available and reviews maybe?