Scarlett Grim

April 20, 2024

Career Foundry

**Task 6.1: Sourcing Open Data**

**Data Source:**

I chose a Denver Airbnb dataset from the [Inside Airbnb](https://insideairbnb.com/get-the-data/). Given that this is coming from the Airbnb website I feel that this dataset is trustworthy.

The dataset contains information about all Airbnb’s in Denver as of December 29, 2023, where there were 4,971 listings. Airbnb indicates that this data is updated quarterly and allows for me to know that this data is current and relevant.

The data set gives each Airbnb an ID, the URL, scraping information such as a scrape id, when it was scraped and what source it came from. It also details the host id, the link to the Airbnb, the host name, host start date, host location, host response rate, host acceptance rate, number of listings a host has, the neighborhood the property is in, the latitude and longitude location of the property, it details the property in terms of privacy, accommodates, bathrooms, availability, and reviews.

I chose one from Denver, Colorado.

I believe this dataset is relevant for my project because it contains a variety of continuous and categorical data. It has the geographical aspects shown by neighborhood but also the latitude and longitude of the listing. The time variable need is met by dates for first and last review along with the host start date.

**Data Choice:**

I chose this data because it is coming from a company website. Any null or blank cells are likely due to a lack of data rather than error. Additionally, I currently live in Denver where my partner and I are buying property to try and create passive income. Understanding the success factors to an Airbnb in Denver would allow for me to have more knowledge going into future property purchases.

**Data Cleaning:**

Deleted columns:

|  |  |  |
| --- | --- | --- |
| **Variable:** | **Action:** | **Explanation:** |
| Description | Delete Column | There is nothing in this entire column there is no purpose to keeping this column. |
| Name | Delete Column | Not needed for the purpose of the analysis. The information it contains – types of properties and number of bedrooms and bathrooms have their own |
| Scrape\_id | Delete Column | Not needed for the purpose of this analysis each listing already has a unique ID |
| Last\_scraped | Delete Column | Not needed for the purpose of this analysis. The scraping is not knowledge I am looking to analyze |
| Source | Delete Column | Source of scrape is not needed for the purpose of the analysis |
| Listing URL | Delete Column | The link to each Airbnb is not needed for the purpose of this analysis |
| Neighborhood overview | Delete Column | A more succinct neighborhood column is already present therefore this elaborate description is not necessary. |
| Picture\_url | Delete column | For the purpose of the analysis there is no need for pictures of the Airbnb’s |
| Host\_url | Delete Column | This is not necessary for the purpose of this project. Each host has a unique ID |
| Host\_name | Delete column | This could be considered PII and given there is a host\_id there is no need for this information. |
| Host\_about | Delete Column | This information exists in other columns and is otherwise unnecessary |
| Host\_thumbnail\_url | Delete Column | Pictures are not needed for the point of this analysis. |
| Host\_picture\_url | Delete Column | Pictures are not needed for the point of this analysis. |
| Host\_neighborhood | Delete column | Not needed for the purpose of this analysis |
| Host listings\_count | Delete columns | Lack of consistency from host listing and host\_total\_listing – therefore removing both columns |
| Host\_total\_listings | Delete\_columns | Lack of consistency from host listing and host\_total\_listing – therefore removing both columns |
| Host\_verification | Delete column | Another column already holds verification status. |
| Host\_has\_profile\_pic | Delete column | All values are true |
| Host\_identity\_verified | Delete column | All values are true |
| Neighbourhood | Delete column | This column just holds a Denver Colorado United States value. Given this is a data set on Airbnb’s in Denver Colorado this information is redundant. More clear neighborhood information exists in the neighbourhood\_cleansed column. |
| Neighborhood\_group\_cleansed | Delete column | The entire column is empty |
| Bathrooms | Delete column | The entire column is empty amd information lies in bathroom\_text column |
| Bedrooms | Delete column | Entire column is empty information lies in ‘beds’ column |
| Amenities | Delete column | Column is empty |
| Minimum\_minimum\_nights | Delete column | This is redundant information |
| Maximum\_minimum\_nights | Delete column | This is redundant information |
| Minimum\_maximum\_nights | Delte columns | This is redundant information |
| Maximum\_maximum nights | Delete column | This is redundant information |
| Minimum\_nights\_avg\_ntm | Delete column | Going to do my own calculations |
| Maximum\_nights\_avg\_ntm | Delete column | Going to do my own calculations |
| Calendar update | Delete column | Empty column |
| Has\_availability | Delte column | All values are true |
| Calendar\_last\_scraped | Delete column | Not needed for the purpose of the analysis |
| License | Delete column | Not needed for the purpose of this analysis |
| Calculated\_host\_listings\_count | Delete column | Redundant information |
| Calculated\_host\_listings\_count\_entire\_homes | Delete column | Redundant information |
| Calculated\_host\_listings\_count\_private\_rooms | Delete column | Redundant information |
| Calculated\_host\_listings\_count\_shared\_rooms | Delete column | Redundant information |

**Further Cleaning:**

**Mixed data types:**

|  |  |
| --- | --- |
| **Variable** | **Resolution** |
| Host\_location | Too many null values - deleted the column |
| Host\_response\_time | Too many null values – deleted the column |
| Host\_response\_rate | Too many null values – deleted the column |
| Host\_acceptance\_rate | Too many null values – deleted the column |
| Host\_is\_super\_host | Only 7 nulls – removing null values |
| Bathrooms\_text | Only 2 nulls – removing nulls |
| Price | 82 null values but this is useful information so I am removing the nulls |
| First\_review | Changed to datetime and null values are now NaT |
| Last\_review | Changed to datetime and null values are now NaT |

Missing data

|  |  |
| --- | --- |
| **Variable** | **Resolution** |
| Beds | Removed the 63 missing values |
| Review\_scores\_rating | All null values were changed to 0 since there are no values of 0 in this column. So, the value 0 represents a lack of reviews not a review score of 0. |
| Review\_scores\_accuracy | All null values were changed to 0 since there are no values of 0 in this column. So, the value 0 represents a lack of reviews not a review score of 0. |
| Review\_scores\_cleanliness | All null values were changed to 0 since there are no values of 0 in this column. So, the value 0 represents a lack of reviews not a review score of 0. |
| Review\_scores\_checkin | All null values were changed to 0 since there are no values of 0 in this column. So, the value 0 represents a lack of reviews not a review score of 0. |
| Review\_scores\_communication | All null values were changed to 0 since there are no values of 0 in this column. So the value 0 represents a lack of reviews not a review score of 0. |
| Review\_scores\_location | All null values were changed to 0 since there are no values of 0 in this column. So, the value 0 represents a lack of reviews not a review score of 0. |
| Review\_scores\_value | All null values were changed to 0 since there are no values of 0 in this column. So, the value 0 represents a lack of reviews not a review score of 0. |
| Reviews\_per\_month | All null values were changed to 0 since there are no values of 0 in this column. So, the value 0 represents a lack of reviews not a review score of 0. |

Renaming Columns

|  |  |
| --- | --- |
| **Variable** | **New name** |
| Neighbourhood\_cleansed | Location |
| accommodates | Guest\_count |
| Bathrooms\_text | bathrooms |

**Data Profile:**

Prior to data cleaning there were 75 variables and 4971 rows.

After removing unnecessary columns and cleaning the data set there are 33 columns and 4817 rows

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Time-variant/Time invariant** | **Structured/unstructured** | **Qualitative/quantitative** | **Qualitative: Nominal/ordinal**  **Quantitative: discrete or continuous** |
| Id | Time-invariant | Structured | Qualitative | Nominal |
| Host\_id | Time-invariant | Structured | Qualitative | Nominal |
| Host\_since | Time-variant | Structured | Quantitative | Discrete |
| Host\_is\_superhost | Time-invariant | Structured | Qualitative | Nominal |
| Location | Time-invariant | Structured | Qualitative | Nominal |
| Latitude | Time-invariant | Structured | Quantitative | Continuous |
| Longitude | Time-invariant | Structured | Quantitative | Continuous |
| Property\_type | Time-invariant | Structured | Qualitative | Nominal |
| Room\_type | Time-invariant | Structured | Qualitative | Nominal |
| Guest\_count | Time-invariant | Structured | Quantitative | Discrete |
| Bathrooms | Time-invariant | Structured | Quantitative | Discrete |
| Beds | Time-invariant | Structured | Quantitative | Discrete |
| Price | Time-invariant | Structured | Quantitative | Continuous |
| Minimum\_nights | Time-variant | Structured | Quantitative | Discrete |
| Maximum\_nights | Time-variant | Structured | Quantitative | Discrete |
| Availability\_30 | Time-variant | Structured | Quantitative | Discrete |
| Availability\_60 | Time-variant | Structured | Quantitative | Discrete |
| Availaibility\_90 | Time-variant | Structured | Quantitative | Discrete |
| Availability\_365 | Time-variant | Structured | Quantitative | Discrete |
| Number\_of\_reviews | Time-variant | Structured | Quantitative | Discrete |
| Number\_of\_reviews\_ltm | Time-variant | Structured | Quantitative | Discrete |
| Number\_of\_reviews\_l30d | Time-variant | Structured | Quantitative | Discrete |
| First\_review | Time-variant | Structured | Quantitative | Discrete |
| Last\_review | Time-variant | Structured | Quantitative | Discrete |
| Review\_scores\_rating | Time-variant | Structured | Quantitative | Continuous |
| Review\_scores\_accuracy | Time-variant | Structured | Quantitative | Continuous |
| Review\_scores\_cleanliness | Time-variant | Structured | Quantitative | Continuous |
| Review\_scores\_checkin | Time-variant | Structured | Quantitative | Continuous |
| Review\_scores\_communication | Time-variant | Structured | Quantitative | Continuous |
| Review\_scores\_location | Time-variant | Structured | Quantitative | Continuous |
| Review\_scores\_value | Time-variant | Structured | Quantitative | Continuous |
| Instant\_bookable | Time-invariant | Structured | Quantitative | Continuous |
| Reviews\_per\_month | Time-variant | Structured | Quantitative | Continuous |

**Descriptive Statistics:**

A screenshot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generated

**Data limitations and ethics:**

After my initial look at the data set, I notice there is no information prior to around 2009. I think that any data prior than this would be helpful for seeing the change overtime. I am worried about the 720 null review values, but I think this is likely due to a lack of reviews not to a lack of values. I think with a flag column and a subset that this could still be very useful information. The website ‘Inside Airbnb’ is made to allow the data sets to be public and for free and the data is updated quarterly so I believe that this information is accurate as it can be while accounting for human error. This transparency reduces the chance of bias.

**Defining questions:**

What types of properties get booked more in the similar price range?

What kind of properties make up the majority of listings and which properties get booked the most?

Are certain neighborhoods in the city more booked than others? How does the spread of Airbnb vary by neighborhood?

What is the average or most popular size of groups staying in Airbnb’s around Denver? How does this compare to the number of bedrooms in different Airbnb’s.

What is the most popular rental duration in Denver? How does this relate to number of guests and size of Airbnb?

What types of listings are not getting booked?

How does being a super host influence the rate that listings getting booked?

Does price per night affect how often a listing is booked?

The highest review is a 5, do reviews lower thana 3 influence the popularity of the listing?