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. 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

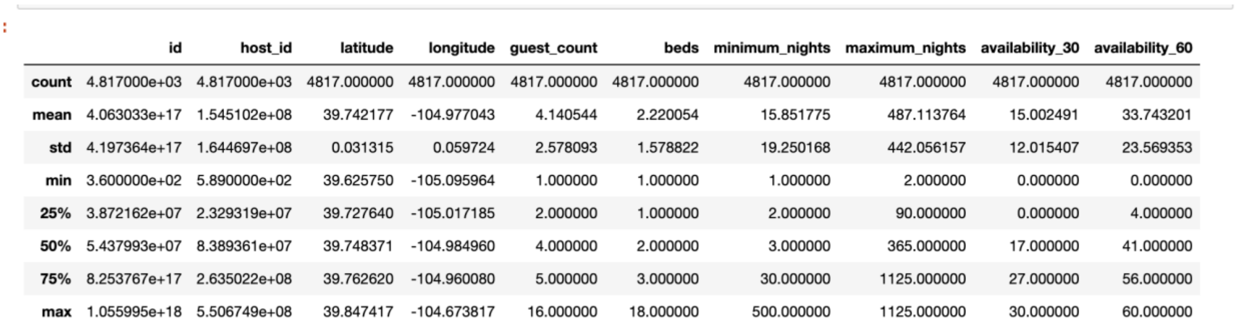
Data Profile:

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

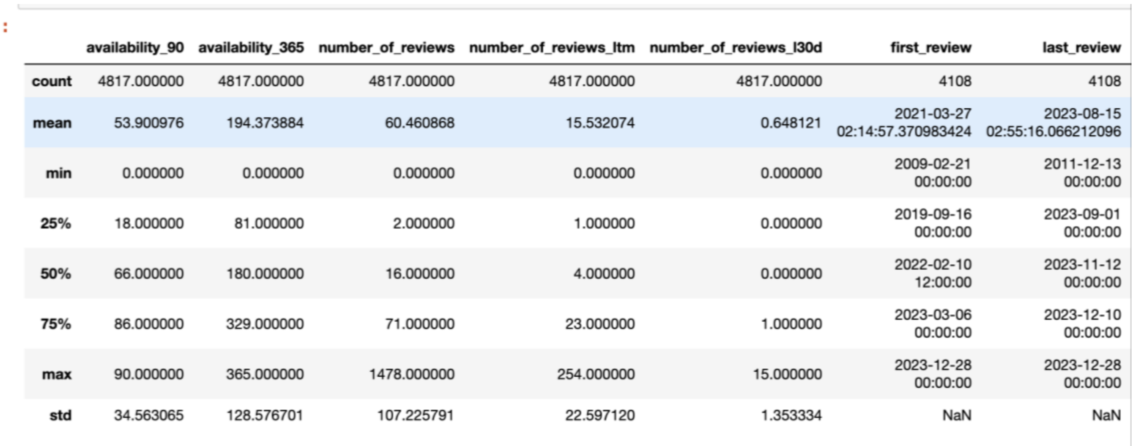
Variable	Time-variant/Time invariant	Structured/unstructured	Qualitative/quantitative	Qualitative: Nominal/ordinal Quantitative: discrete or continuous
Id	Time-invariant	Structured	Qualitative	Nominal
Host_id	Time-invariant	Structured	Qualitative	Nominal
Host_since	Time-variant	Structured	Quantitative	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

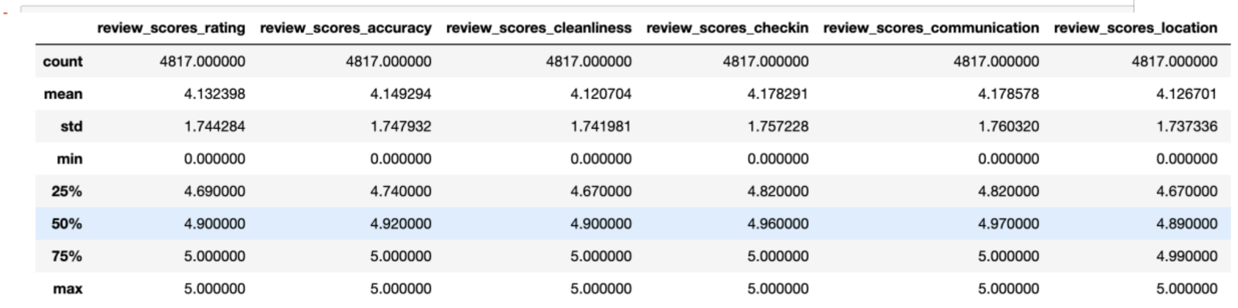
Descriptive Statistics:



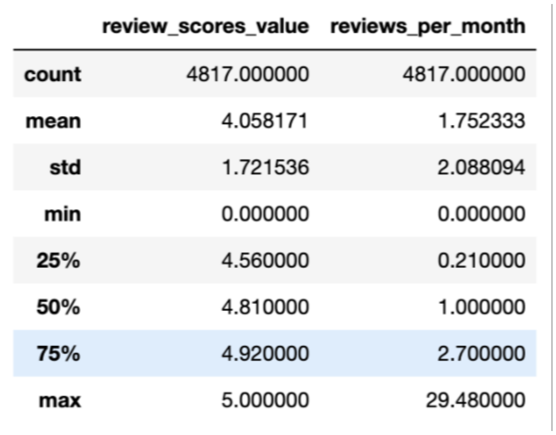
	id	host_id	latitude	longitude	guest_count	beds	minimum_nights	maximum_nights	availability_30	availability_60
count	4.817000e+03	4.817000e+03	4817.000000	4817.000000	4817.000000	4817.000000	4817.000000	4817.000000	4817.000000	4817.000000
mean	4.063033e+17	1.545102e+08	39.742177	-104.977043	4.140544	2.220054	15.851775	487.113764	15.002491	33.743201
std	4.197364e+17	1.644697e+08	0.031315	0.059724	2.578093	1.578822	19.250168	442.056157	12.015407	23.569353
min	3.600000e+02	5.890000e+02	39.625750	-105.095964	1.000000	1.000000	1.000000	2.000000	0.000000	0.000000
25%	3.872162e+07	2.329319e+07	39.727640	-105.017185	2.000000	1.000000	2.000000	90.000000	0.000000	4.000000
50%	5.437993e+07	8.389361e+07	39.748371	-104.984960	4.000000	2.000000	3.000000	365.000000	17.000000	41.000000
75%	8.253767e+17	2.635022e+08	39.762620	-104.960080	5.000000	3.000000	30.000000	1125.000000	27.000000	56.000000
max	1.055995e+18	5.506749e+08	39.847417	-104.673817	16.000000	18.000000	500.000000	1125.000000	30.000000	60.000000



	availability_90	availability_365	number_of_reviews	number_of_reviews_ltm	number_of_reviews_l30d	first_review	last_review
count	4817.000000	4817.000000	4817.000000	4817.000000	4817.000000	4108	4108
mean	53.900976	194.373884	60.460868	15.532074	0.648121	2021-03-27 02:14:57.370983424	2023-08-15 02:55:16.066212096
min	0.000000	0.000000	0.000000	0.000000	0.000000	2009-02-21 00:00:00	2011-12-13 00:00:00
25%	18.000000	81.000000	2.000000	1.000000	0.000000	2019-09-16 00:00:00	2023-09-01 00:00:00
50%	66.000000	180.000000	16.000000	4.000000	0.000000	2022-02-10 12:00:00	2023-11-12 00:00:00
75%	86.000000	329.000000	71.000000	23.000000	1.000000	2023-03-06 00:00:00	2023-12-10 00:00:00
max	90.000000	365.000000	1478.000000	254.000000	15.000000	2023-12-28 00:00:00	2023-12-28 00:00:00
std	34.563065	128.576701	107.225791	22.597120	1.353334	NaN	NaN



	review_scores_rating	review_scores_accuracy	review_scores_cleanliness	review_scores_checkin	review_scores_communication	review_scores_location
count	4817.000000	4817.000000	4817.000000	4817.000000	4817.000000	4817.000000
mean	4.132398	4.149294	4.120704	4.178291	4.178578	4.126701
std	1.744284	1.747932	1.741981	1.757228	1.760320	1.737336
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	4.690000	4.740000	4.670000	4.820000	4.820000	4.670000
50%	4.900000	4.920000	4.900000	4.960000	4.970000	4.890000
75%	5.000000	5.000000	5.000000	5.000000	5.000000	4.990000
max	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000



	review_scores_value	reviews_per_month
count	4817.000000	4817.000000
mean	4.058171	1.752333
std	1.721536	2.088094
min	0.000000	0.000000
25%	4.560000	0.210000
50%	4.810000	1.000000
75%	4.920000	2.700000
max	5.000000	29.480000

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?