Short-term rentals and home price analysis

Executive summary

Recommendation 1: increase marketing spend for properties in the 78703 zipcode.
Recommendation 2: confirm demand for larger properties through targeted advertising campaign

Table of Contents

Introduction

Touted as the "Live Music Capital of the World", Austin is the state capital of Texas and is well known for its country, blues, and rock music – and performers lining everywhere from the music festivals to the grocery stores. It has many parks and lakes, popular to purveyors of an outdoor lifestyle – hiking, biking, swimming and boating are popular pastimes – especially given that the weather is generally predictable year-round. The University of Texas campus is also primarily located here and serves as a hub for tech start-ups.[1] Given these facts, it is reasonable to suggest that visitors travelling to Austin may come for shorter periods, for instance, around particular events such as South by Southwest.

Short-term rentals of property are increasingly seen as a secondary source of income for individuals in addition to businesses making home stays a viable alternative to hotel stays. Airbnb is a platform that provides homes to rent, acting as a broker between the "host" and "guest" and collects a commission on that deal. One goal Airbnb has is to onboard more diverse renters onto the platform through advertising to gain more revenue from customers whose needs are not being met. Additionally, Airbnb could increase revenue by better servicing the needs of the host and providing education about what makes a better rental property in the eyes of the guest customers.

Analysis

Ensuring the raw data remained untouched for versioning purposes, data were staged, formatted and cleaned for analysis. For all detailed methods, see Appendix 2: Methods.

Numeric clean data were run through a correlation matrix by table to identify some initial trends to investigate further. Figure 1 shows the correlation matrix for Airbnb listing data. Some interesting observations are included below, while other findings can be found in the appendix.

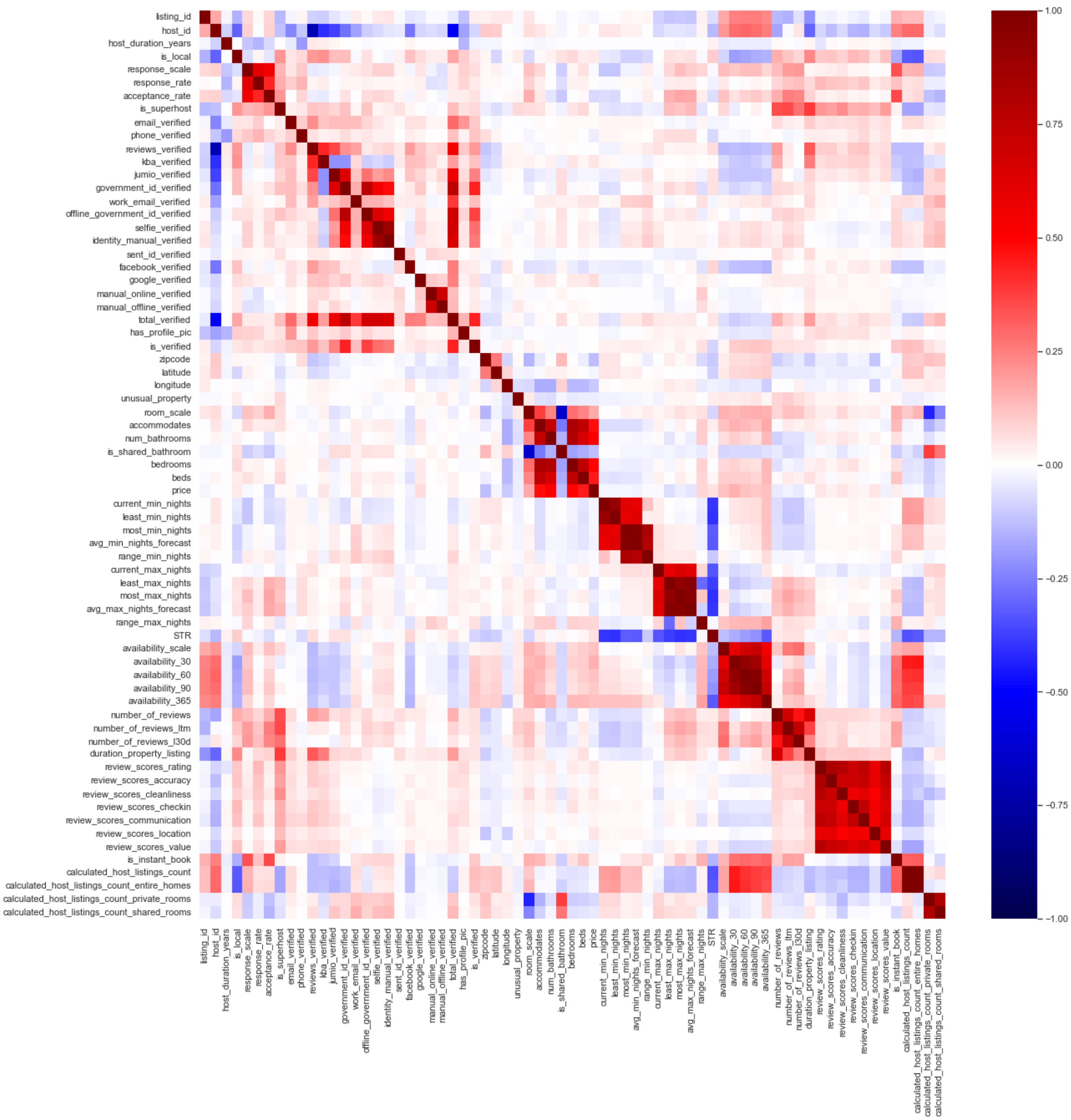


Figure 1: Airbnb listing correlation matrix

1) Shared properties are more likely to have shared bathroom arrangements, whereas entire homes are associated with more bedrooms and bathrooms. These factors may contribute to the higher price requested by hosts of entire properties.
2) Short term rentals are negatively associated with hosts who have a higher listing count, especially of entire properties; therefore, hosts with more listings have fewer properties that can be booked for less than 30 days. I suspect that individuals with many listings will be driven to try to host guests looking for a longer stay, since this is likely to increase revenue and reduce the costs associated with turning around a property for the next guest, especially when done at scale.
3) Short term rentals are also negatively associated with availability over the next 365 days, suggesting that there is less availability for short-term stays in this dataset.
4) Additionally, hosts who have a high listings count, especially of entire properties, have more availability over the next 365 days than hosts with fewer properties, and are more likely to have the instant book feature, probably driven by the need to fill the properties with less interaction with each individual guest.

The property-level information gained from this correlation matrix is perhaps unsurprising, but indicates a strong correlation between the size of the property (using the number of bedrooms as a proxy) and price of the listing. Short term rentals appear to be preferred by hosts with fewer listings; perhaps as a consequence of hosts with more listing stock having the luxury of being able to wait for the ideal guest.

A similar exercise was conducted for Zillow data (

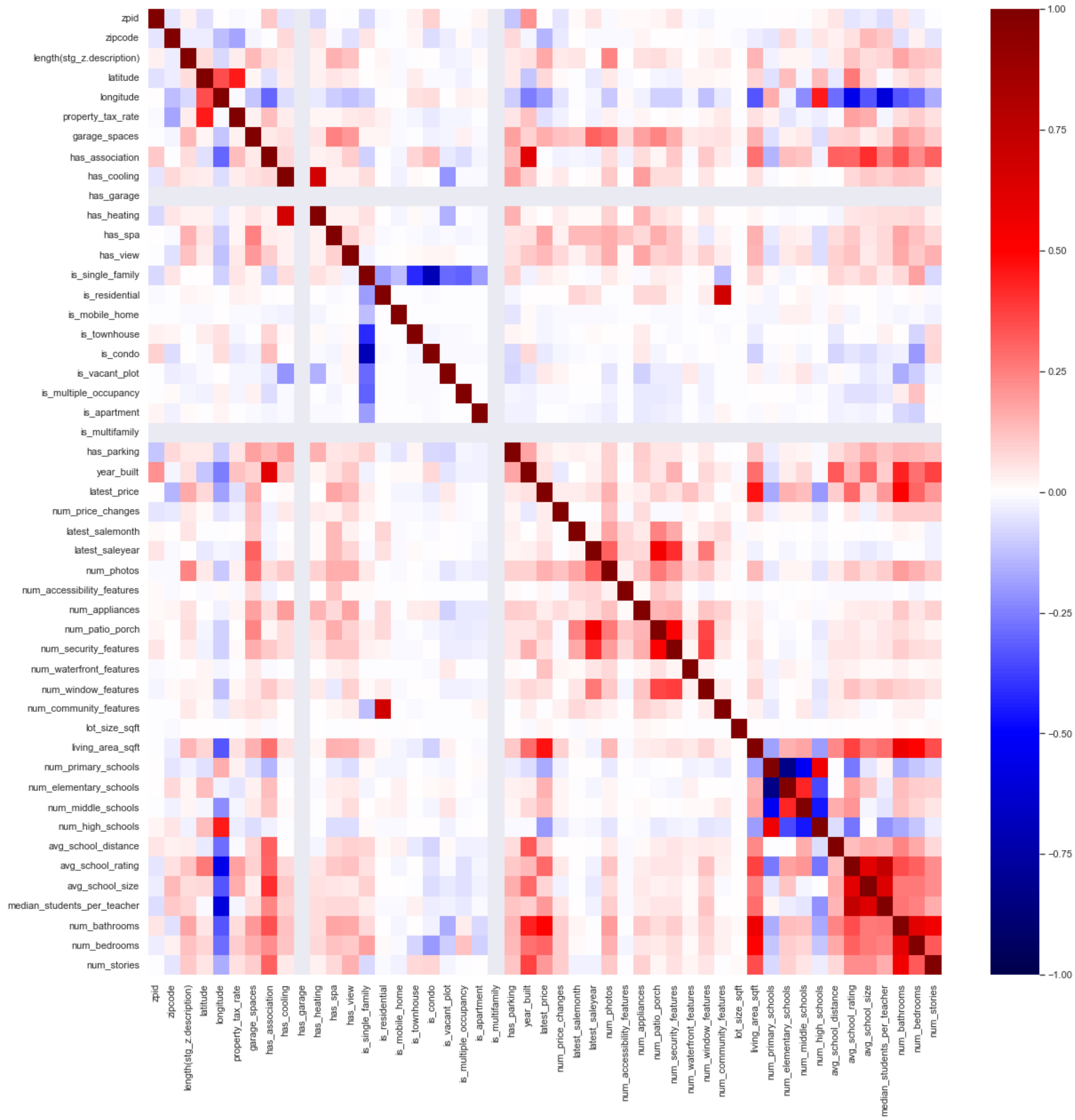


Figure 2), with the following interesting observations:

1) There is a negative correlation between location (longitude) and sqft of property, and school metrics (median students per teacher, average school size and rating, and distance of the school from the property. This suggests an east to west divide across the city; properties in the east of the city are larger, but have fewer high schools in the vicinity, larger distances between the property and the school, higher school ratings but also a higher number of students per teacher. These factors could all influence the desirability of purchasing a home in these regions. Additionally, properties on the east

are associated with larger properties, both in terms of square footage and number of bedrooms.

2) Newer properties have more square footage and more bedrooms in the property than older properties. This could act as a proxy indicator for what buyers look for when purchasing property.
3) Price seems to be driven by the sqft of the house, in addition to the number of bedrooms and the rating of the local schools. Larger properties are associated with more bedrooms and bathrooms; however, these properties have fewer primary and high schools in the vicinity.

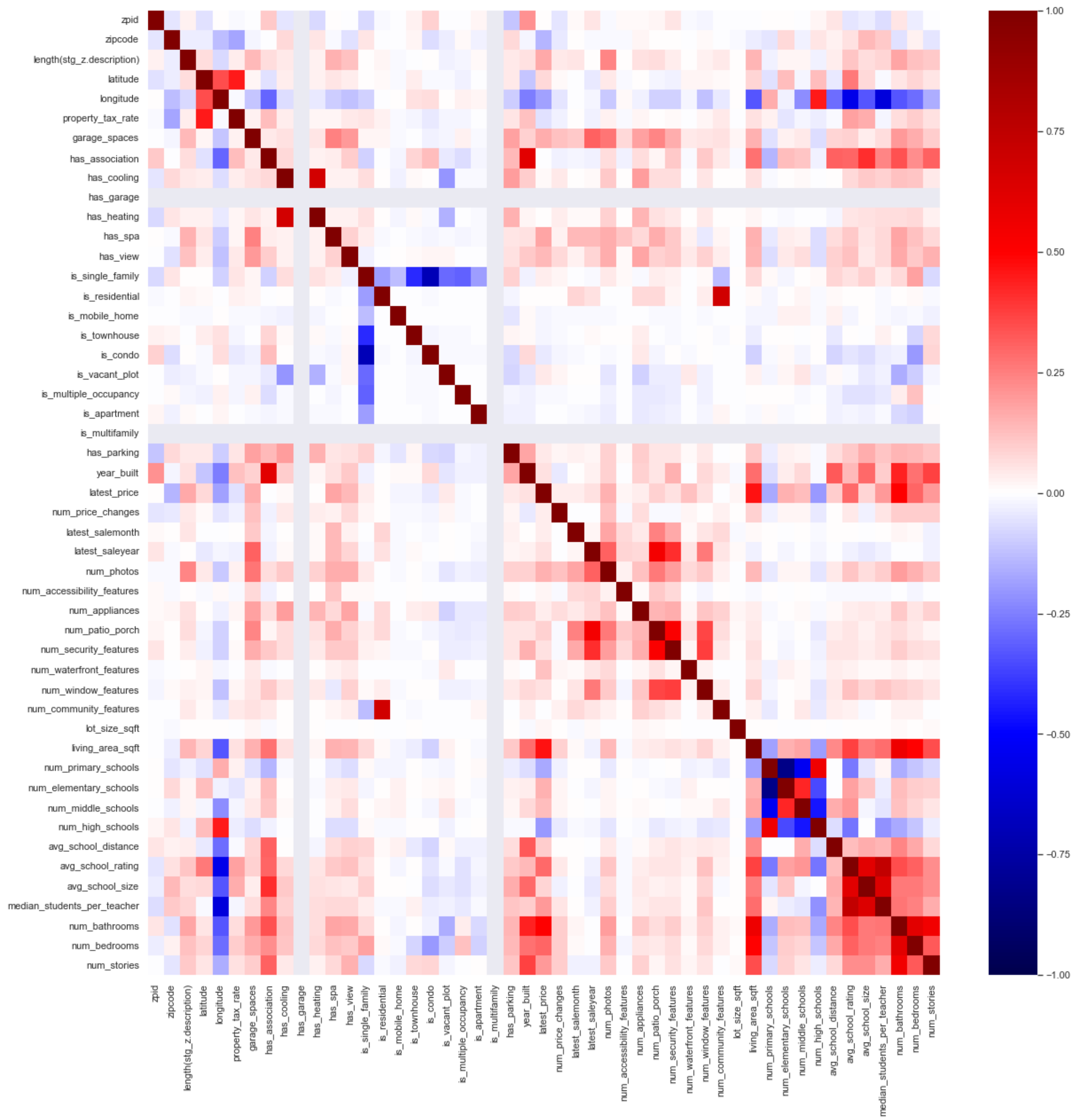


Figure 2: Zillow listing correlation matrix

Property analysis

In the first instance, we want to discover differences between successful listings on Airbnb and what drives price variability in relation to short term rentals, which are defined as fewer than 30 days[2]. Figure 3 shows the price changes by room type. As shown in part B, entire properties drive the listings on Airbnb, and make up 82.7% of the properties that have had at least one stay[3]. Many more properties across the dataset fall in the longer-term stay category, and could indicate a demand for listings that permit short term rentals. The short-term rentals, by contrast, make up only 13.2% of the dataset (part D). Parts A and C of the figure show the distribution of price data in the dataset and while they do not infer causality, imply that most users of the platform stay in entire properties over other types of room even though prices are lower in other types of room. Interestingly, hotel room data show more expensive longer-term stays than short term stays, which is in contrast to shared spaces (shared rooms and private rooms in shared properties); however, there are too few data (10 listings) to derive meaningful insights from hotel room data. There is a small increase in the listing price of a long-term rental for an entire property but this appears to be too small to be meaningful.

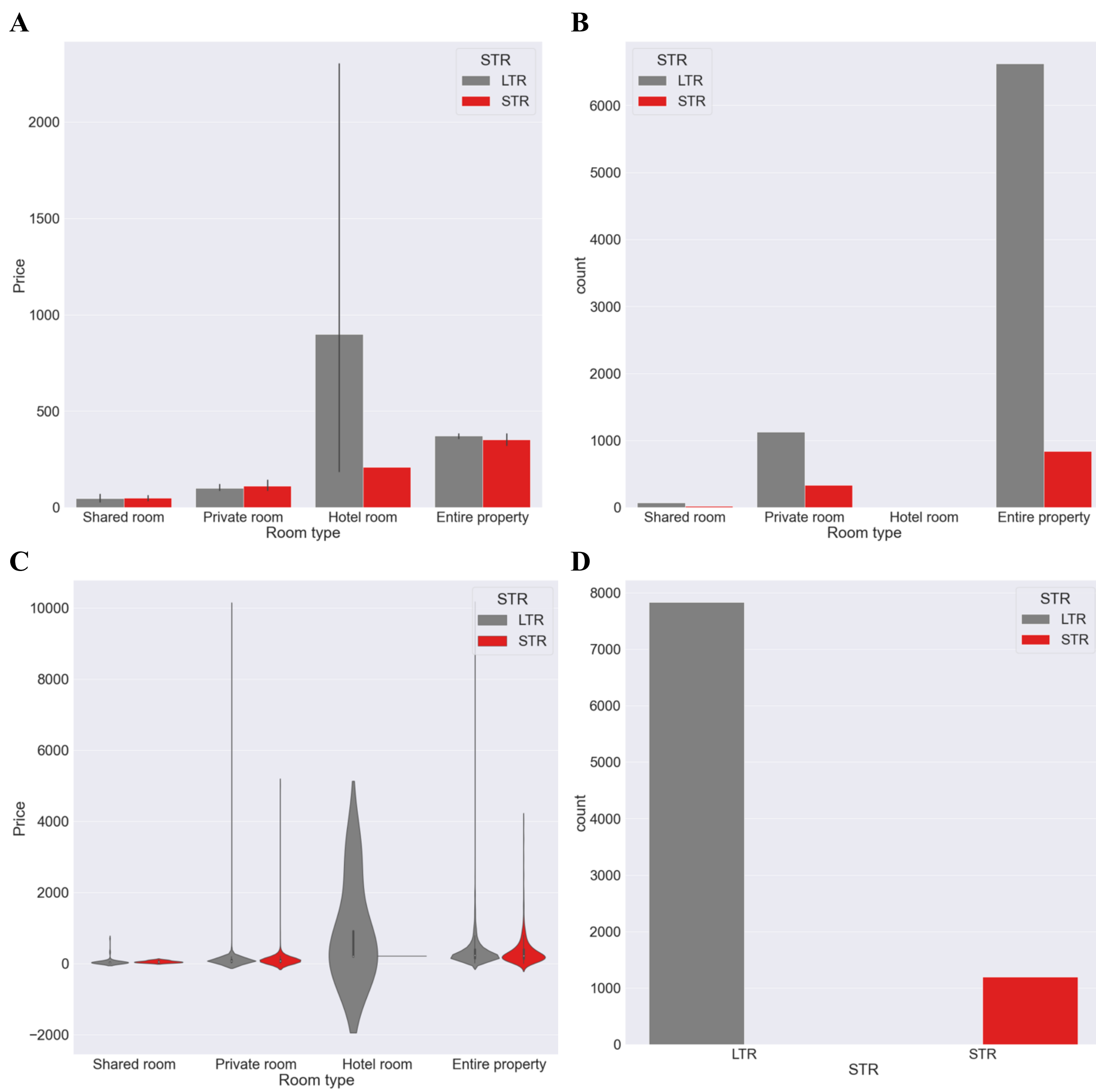


Figure 3: Airbnb segmentation analysis with short-term rental (red). A) Comparison of price across different property types. B) Number of listings according to property type. C) Distribution of price by property type. D) Count of short-term and long term listings in the Airbnb dataset

To find more meaningful intuition with regard to whole property listings and square footage, further analyses of these data only used entire property listings from the Airbnb dataset. In Figure 4, Airbnb listing price is segmented by number of rooms in the property, showing little difference between short-term rentals and longer-term rentals except when observing properties with 5 or more bedrooms. These properties can command a 15% increase in the listing price as a short-term rental than the longer-term counterpart. This finding is amplified when considering the price per room, shown in part B. The price of the property per room is maintained across Studio, 1-, 3-, and 4-bedroom properties, with an increase in price per room for properties with

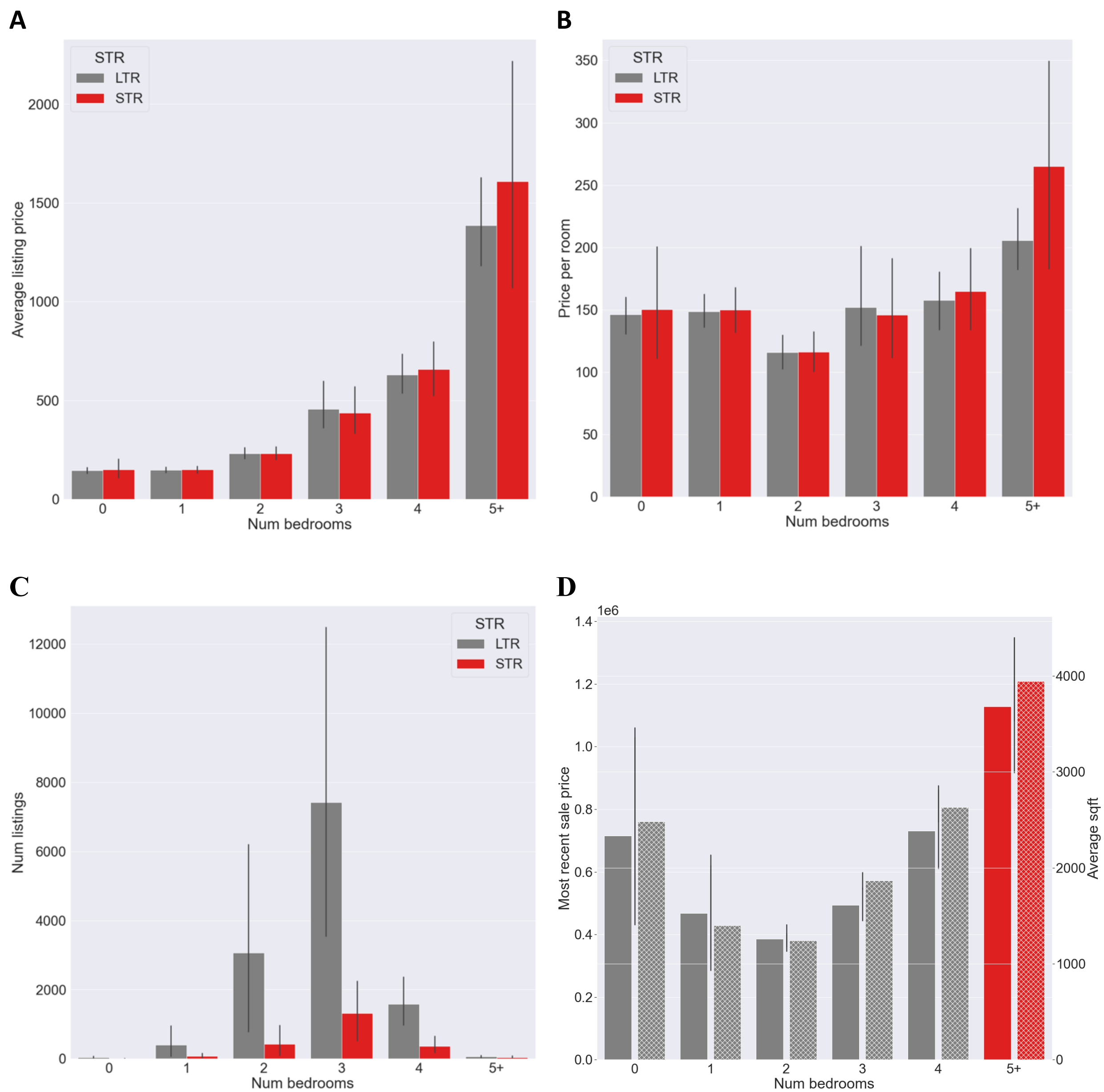
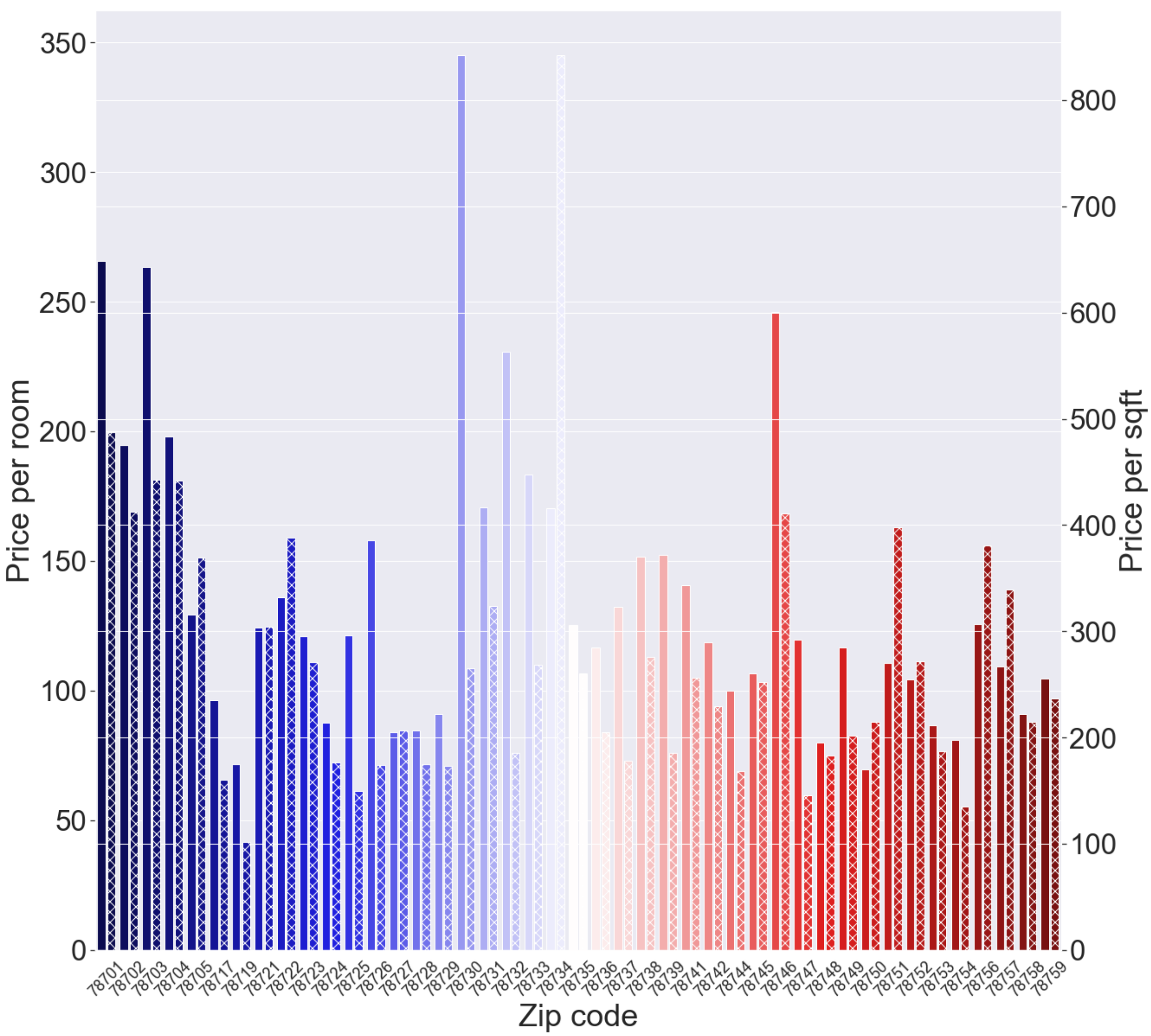


Figure 4: Segmentation Airbnb data by number of bedrooms in a property. A) Comparison of price across number of rooms in a property. B) Listing price per room by property size. C) Number of listings by number of bedrooms in the property. D) Airbnb price per room and Zillow price per square foot approximate each other for the purposes of joining these data

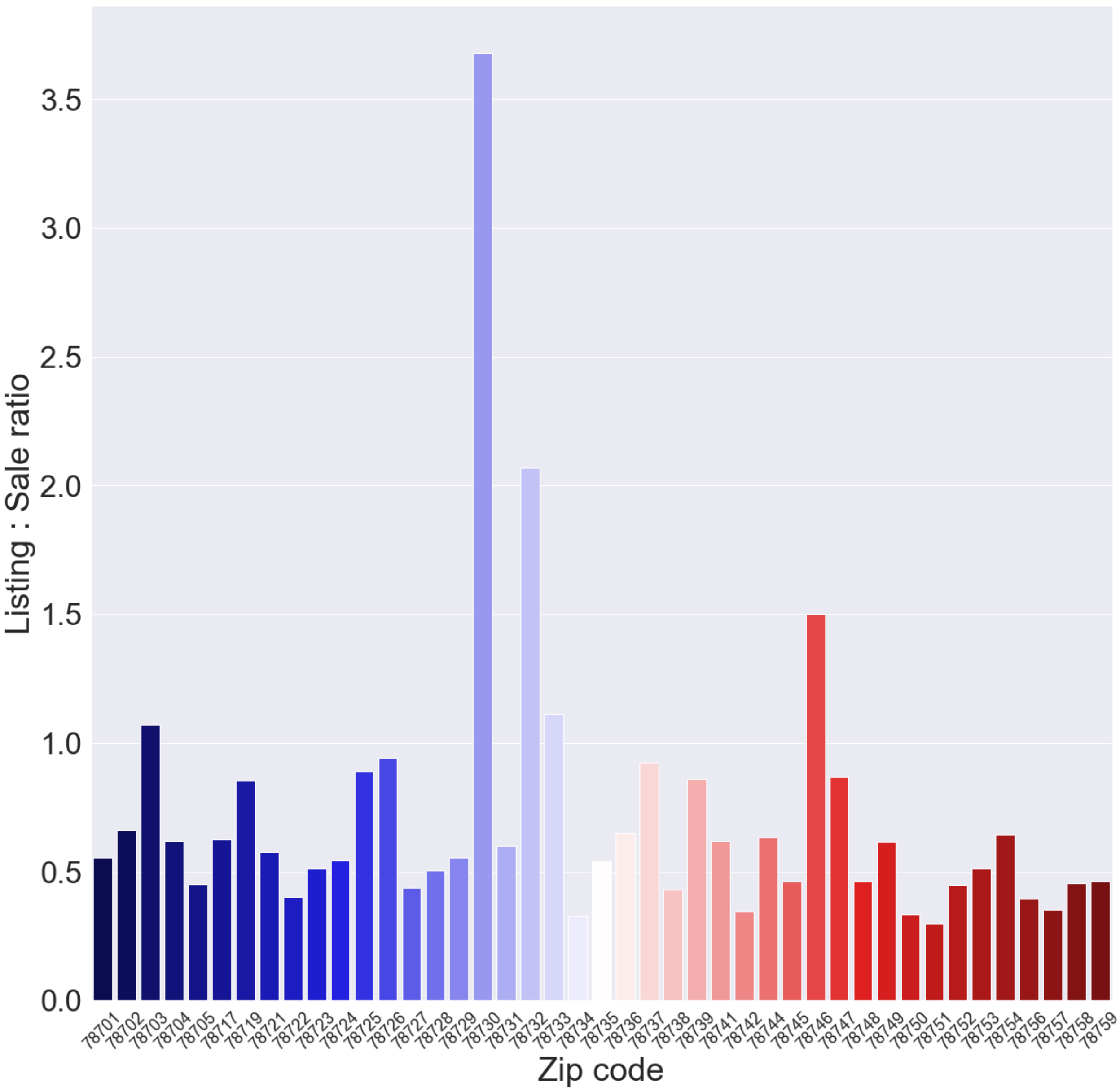
5 or more bedrooms, especially in short-term rentals. This could be indicative of supply: there may be fewer properties with 5 or more bedrooms, and hosts of these properties are able to list their properties at a premium. This finding is confirmed with the plot in part C, which shows the number of properties segmented by size and short-term rentals. Part D, showing the number of bedrooms of a property and the most recent listed sale price and the size of the property in square feet, shows a close association between these factors, and indicates that we may be able to use price per square foot to compare with price per room in the Airbnb dataset.

The final part of the property-level segmentation analysis surrounds where in the city Airbnb hosts list their properties, the availability of properties for sale, and the cost of purchasing properties in the area. Comparing the listing price with the average sale price in each zip code provides some interesting initial insight (Figure 5). Part A compares the price per room with

A

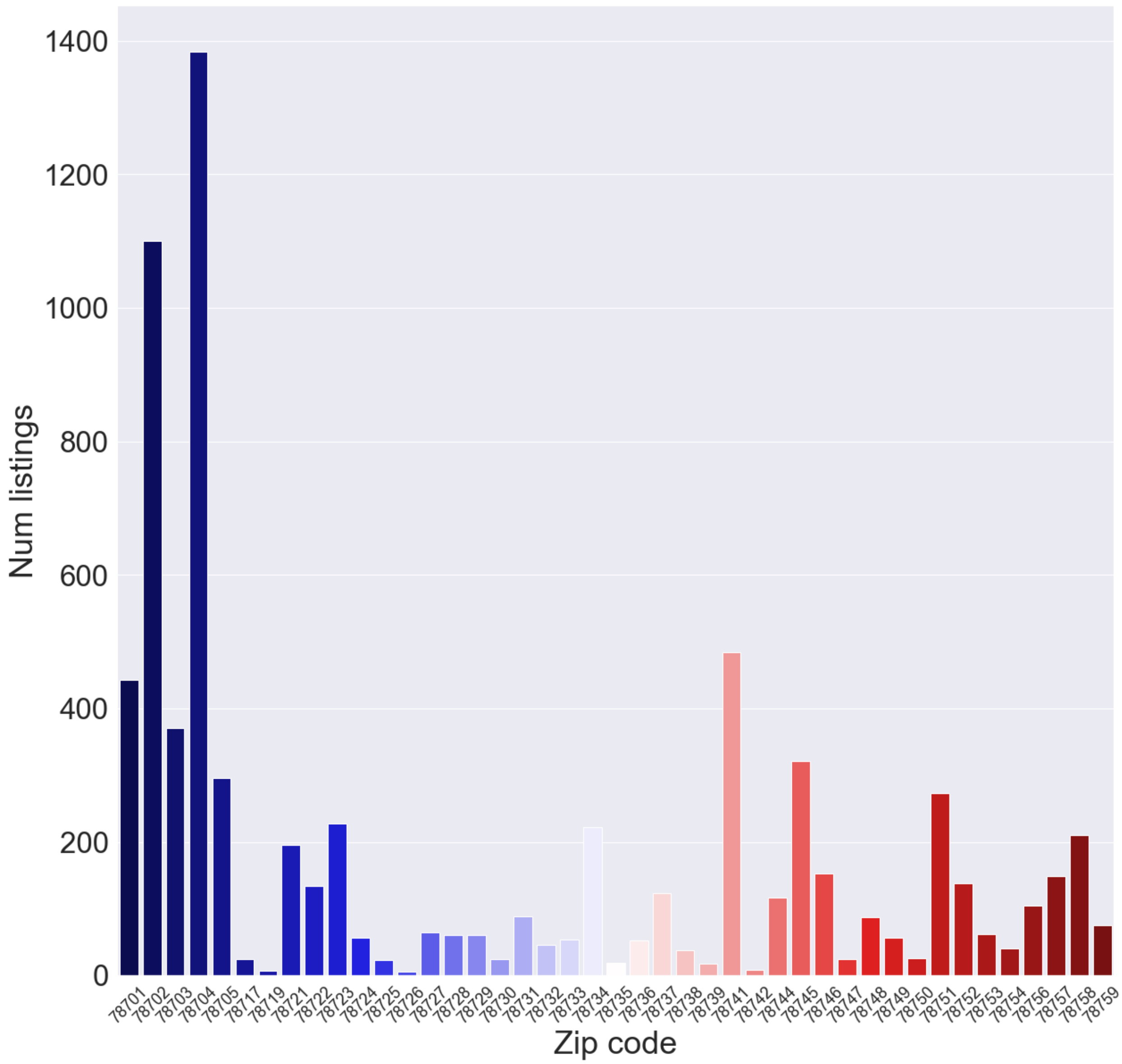


B



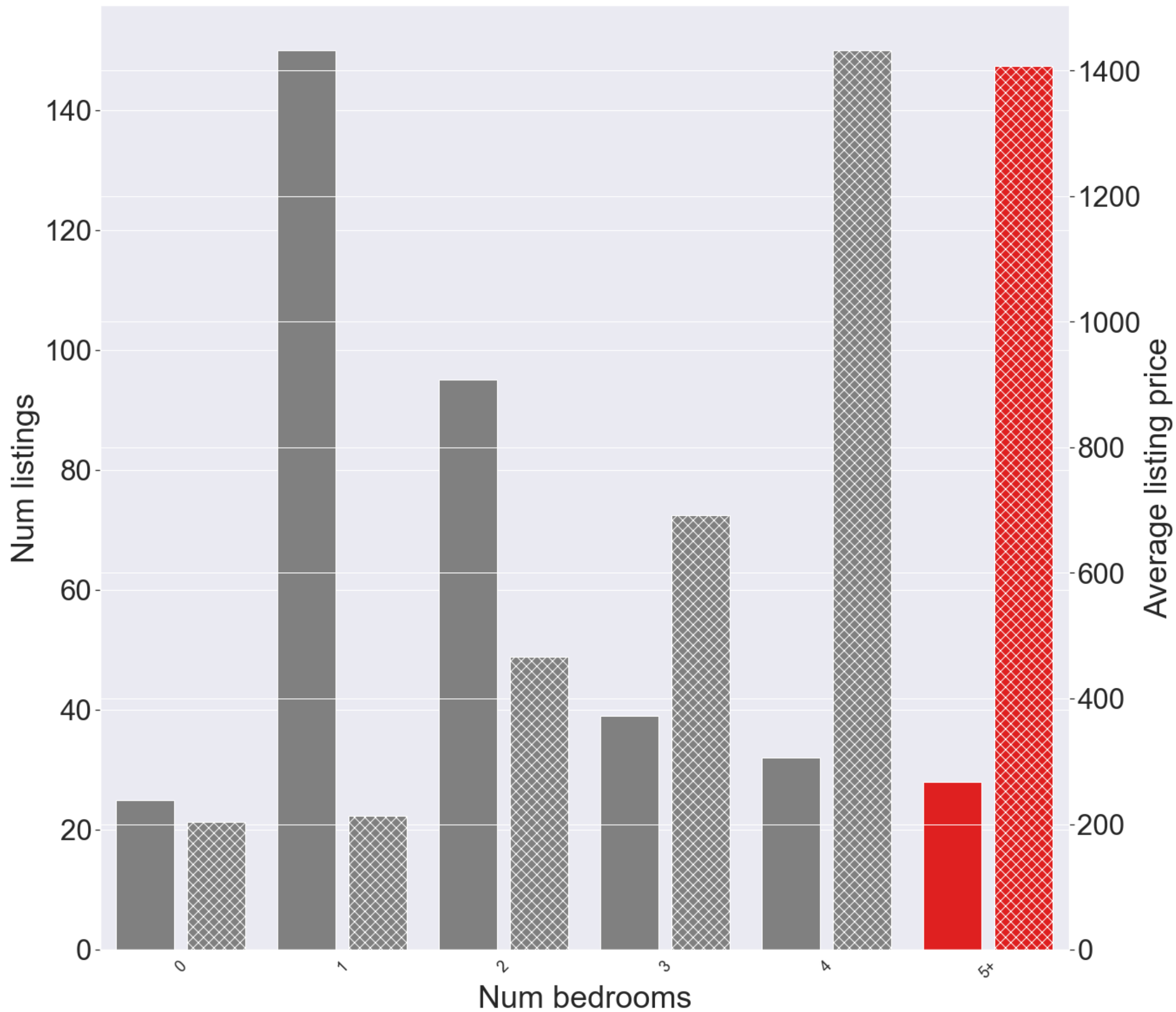
Listing : Sale ratio
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
78701
78702
78703
78704
78705
78717
78719
78721
78722
78723
78724
78725
78726
78727
78728
78729
78730
78731
78732
78733
78734
78735
78736
78737
78738
78739
78741
78742
78744
78745
78746
78747
78748
78749
78750
78751
78752
78753
78754
78756
78757
78758
78759
Zip code

C



Num listings
1400
1200
1000
800
600
400
200
0
78701
78702
78703
78704
78705
78717
78719
78721
78722
78723
78724
78725
78726
78727
78728
78729
78730
78731
78732
78733
78734
78735
78736
78737
78738
78739
78741
78742
78744
78745
78746
78747
78748
78749
78750
78751
78752
78753
78754
78756
78757
78758
78759
Zip code

D



Num listings
0
20
40
60
80
100
120
140
Average listing price
0
200
400
600
800
1000
1200
1400
0
1
2
3
4
5+
Num bedrooms

E

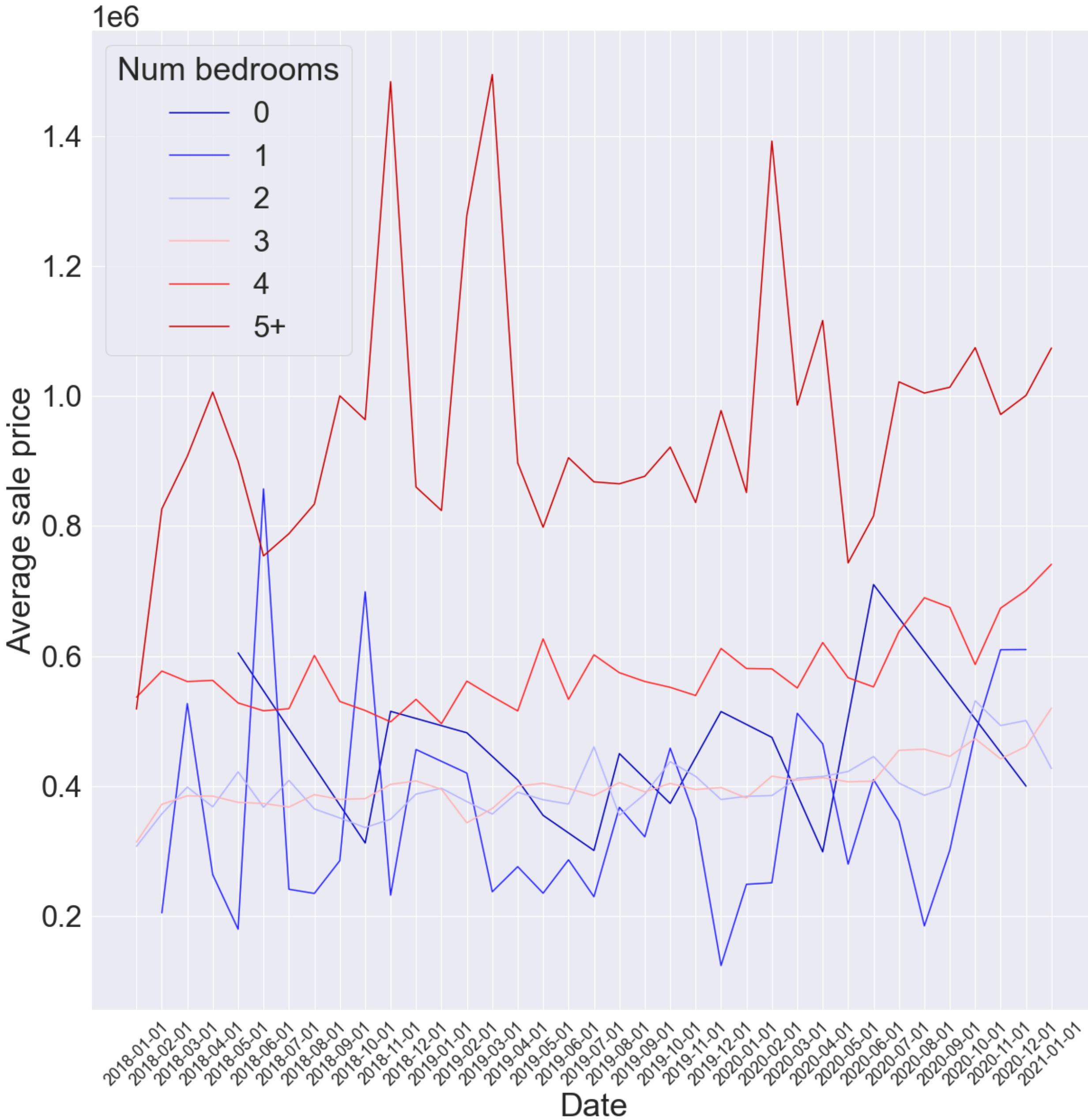


Figure 5: Zipcode analysis of Airbnb and Zillow listing data. A) Price per room and price per square foot segmented by zipcode. B) Derived ratio of rental price / purchase price segmented by zipcode. C) Number of listings by zipcode. D) listing price and number of listings in zipcode 78703 by number of rooms. E) Trends in sale prices of properties in 78703 zipcode.

price per square foot. Adjusting for size provides a reasonable comparison between the listing and sale price of a property, which is generated into a ratio in part B, where lower values indicate a large return on investment compared with the sale price of the property in relation to its size. From this analysis, 5 zipcodes are identified as having the most potential: 78730, 78732, 78746, 78733, and 78703. On counting the number of listings available across zipcodes, we discover that the first four zipcodes have low representation on Airbnb, with fewer than 200 properties in each area; however, in 78703 there are 400 listings, with high listing prices for 4- and 5-

bedroom properties (part D), where the supply is low. Trends in sale price in this zipcode over time indicate a largely flat market for 2- and 3- bedroom properties, and large fluctuations in 0- and 1-bedroom properties, which are difficult to interpret. However, a small upward trend is seen in 4- and 5- bedroom properties in this area of Austin.

Conclusions

Our goal for this exercise was to derive some insights into the short-term rentals in the Airbnb listing data and listing price. In analysing the given datasets, we have reached the following conclusions:

First, properties in zipcode 78703 in Austin, Texas appear to have potential when viewing the price per room listed in Airbnb and price per square foot in Zillow, indicating that investors in property in Austin may be interested in seeking out properties in this region for a high return on their investment. On further investigation, this zipcode is home to Tarrytown, a neighbourhood of Austin that is by the river, an art museum, golf courses, and has easy access to Lake Austin and the trails and national parks, while still having proximity to the buzz of downtown Austin. For Airbnb, increasing marketing and awareness campaigns will drive the number of hosts onboarding to the platform and increase supply of properties in this region.

Recommendation 1: increase marketing spend for properties in the 78703 zipcode.
For example, a marketing campaign targeting users residing in this region could make use of a breakeven analysis, incorporating mortgage data and providing information around how much a property could be rented for in a calendar year to cover repayments and minimize the financial risk of property ownership to the individual.

Second, there is demand for properties with 5 or more bedrooms. This is inferred from there being a low availability and a high price per bedroom. There could be several reasons for this finding. One suggestion is that larger houses are easier to split amongst friends; however, we do not have access to stay data, although this would be an interesting addition to the dataset for future analyses. There is also the option that these properties have other amenities that make them desirable, which is something that could be looked into further.

Recommendation 2: confirm demand for larger properties through targeted advertising campaign
We can identify through A/B testing an advertisement campaign. Advertisements showing larger properties, and smaller properties with desirable ameneties (eg, a swimming pool) could be run to identify the causality of the desire for larger properties.

Next steps

- Discover properties that are not doing well that should be. How can we educate hosts to have their place rented more (so that Airbnb can make profit)? Educating hosts on listing for the right price (taking into account amenities, zipcode, and other factors), and identifying both when a property is being listed for too much or too little is important for success on the Airbnb platform.

- Due to the lack of granular data in the authorized rentals dataset, we were not able to explore licensing information in this analysis. Does adding a license for short term rentals affect renter decisions?
- Although Airbnb will never be sure of the way a consumer will respond to changes in pricing, developing an elastic pricing model to advise hosts on pricing in times of increased or reduced supply of particular types of property would allow experimentation around raising prices on entire properties while maintaining customer conversion. This model will better inform hosts on their pricing strategy and could result in their listings being likely to get booked.
- Applying a seasonal model such as (SARIMA) – uses differences in seasons to remove seasonal effects. For instance annual music festivals such as South by South West can drive annual fluctuations in supply. In a SARIMA model, recommendations could be made as to increasing campaign spend in the months leading to an event.

Appendices

Appendix 1: Assumptions and Caveats

1) Given what we know about Austin, Texas, my first assumption is that most AirBnB rentals are for tourism rather than business.
2) For the purposes of analyses presented here, short term rental properties are defined as those that only permit short term rentals (ie, the calendar function only permits stays that are between 1 and 30 days long). LTR are all other rentals. This was decided as a strategy because removing all ambiguous data (ie, properties that allowed both long and short term stays) would remove approximately 2/3 of the dataset.
3) Properties with no stays have been excluded because they add noise to the dataset when considering properties that are good investment opportunities.
4) Without stay data it is difficult to glean some of the more nuanced information, such as how much guests are willing to spend per person. I have used price per room as a proxy for this information. Using the accommodates column would have been an alternative, but the needs of groups of guests may change based on their circumstances whereas the number of rooms in the property does not, although the reasons for booking those properties might.

Appendix 2: Methods

Data Cleaning

First, all tables were given types and keys in the staging data tables, labelled stg_*table name*. All tables and columns were formatted into snakecase according to SQL formatting best practices. Dates for columns of interest were formatted into 'YYYY-MM-DD' format according to SQLite requirements. ID columns were checked for uniqueness and visual inspections of categorical data to ensure minimal duplication.

Each staging table was then cleaned further. The following schema describe the working assumptions for each view:

cleaned_airbnb

Table column	**Working definition / logic of derived column**
listing_id	[**PRIMARY KEY**] [**ID**] [**NOT NULL**] The identifier for each row in the dataset. No duplicates or nulls were permitted in this column.
name	[**TEXT**] Name of the listing, generated by the host.
length_description	[**NUMERIC**] A count of each character in the long description of the property. Perhaps

	hosts who spend more effort writing about their property generate more revenue.
length_neighbourhood_overview	[**NUMERIC**] A count of each character in the description of the neighbourhood where the host property is located.
length_about	**NUMERIC**] A count of each character in the description of the host.
host_id	[**ID**] [**NOT NULL**] A unique identifier for each host
host_name	[**TEXT**] The name of the host. On visual inspection, it could be useful to separate hosts based on whether they are an individual or a business.
host_since	[**DATE**] The date the host joined Airbnb – not necessarily with the property in the given listing_id row. This is a fluid data type in SQLite, but must be converted to the format YYYY-MM-DD to be identified as a date. *substr(mydate, 7)* || '-' || *substr(mydate, 1,2)* || '-' || *substr(mydate, 4,2)*
host_duration_years	[**NUMERIC**] The number of years the host has been on the platform *DATE('now') – host_since*
is_local	[**NUMERIC**] A derived column from the host_location column in the stg_airbnb_listing data that identifies hosts based in Austin and provides a binary output for easier analysis. Local hosts might be preferred to those from out of the city, and there might be interesting information to glean from hosts from out-of-town (investors).
response_scale	[**NUMERIC**] A derived column from the host_response_time column in the stg_airbnb_listing table. Since the data are ordinal and therefore a numbered scale made sense in this context ranging from 0 (no response) to 4 (within an hour). Null data were interpreted as never receiving a request and were assigned a null value. *Null = No data* *0 = No response* *1 = Within a few days or more* *2 = Within a day* *3 = Within a few hours* *4 = Within an hour*

response_rate	[**NUMERIC**] Host response rate was converted to a decimal from a percentage, with empty strings converted to null values to signify no contact
acceptance_rate	[**NUMERIC**] Host acceptance rate was converted to a decimal from a percentage, with empty strings converted to null values. This column formed part of the logic for filtering out data with no stays.
is_superhost	[**NUMERIC**] The host_is_superhost column was converted into binary from “t” and “f” for correlation analysis. Nulls were included in the false condition because they are not, by definition, true.
host_neighbourhood	[**TEXT**] Enforced null for blank rows. However, there are 503 distinct neighbourhoods, but only 40 listed neighbourhoods in my research. I will come back to this if there is time and it seems worthwhile to force data. For now, sufficient information can be acquired from other columns. *SELECT COUNT (DISTINCT host_neighbourhood) FROM stg_airbnb_listing* ***yields 503***
{***verification method***}_verified	[**NUMERIC**] 15 columns representing the unique contents of host_verifications column. By creating dummy variables, we are able to ascertain if any particular verification method is preferred by Airbnb guests. Defined as 1 if the host has been verified using that method and 0 if not. The columns were developed using a CTE.
total_verified	[**NUMERIC**] Sum of the {***verification method***}_verified columns
has_profile_pic	[**NUMERIC**] ‘t’ and ‘f’ converted to 1 and 0, respectively for correlation analysis. This column was preferred over the columns with url for the host picture, since it also spoke to the contents of the image. The host_picture_URL column in the stg_airbnb_listing column also provided links to placeholder images.
is_verified	[**NUMERIC**] Binary column defining listings that are considered by Airbnb to be verified hosts

zipcode	[**TEXT**] [**FOREIGN KEY**] Column information was incorrectly labelled as neighbourhood_cleansed in the stg_airbnb_listing table. Given the lack of property data, zipcode will be the foreign key for joining aggregated zip-level data between the Airbnb and Zillow listings.
Latitude	[**TEXT**] Unchanged from the original data. One point to note was that other coordinate data were given to 8 decimal places, whereas the stg_airbnb_data only has 5
latitude	[**TEXT**] Unchanged from the original data. One point to note was that other coordinate data were given to 8 decimal places, whereas the stg_airbnb_data only has 5
property_type	[**TEXT**] Type of property (eg, apartment, condo, treehouse). There were some potential duplications
unusual_property	[**NUMERIC**] A derived column based on the previous that identified “unusual properties” such as bus, treehouse, farm as a feature, adding to the idea that the stay at an Airbnb is an experience rather than a hotel. This was developed using a CTE.
room_scale	[**NUMERIC**] A derived feature from the room_type column from the stg_airbnb_listing table that provided ordered categorization from the most shared property to the least: *0 = shared room* *1 = private room in shared property* *2 = hotel room* *3 = entire property*
accommodates	[**NUMERIC**] The total number of people a host property can sleep (in beds or couches or other sleeping arrangements)
num_bathrooms	[**NUMERIC**] A string function was used to split the bathrooms_text column to retrieve only the number of bathrooms: *CAST(substr(bathrooms_text, 1, instr(bathrooms_text, ' ')) AS INTEGER)* The **substring** function permits the search through the text, while the **in string function** here searches for the first instance of a ‘ ‘ and returns the contents before that string.

is_shared_bathroom	[**NUMERIC**] A binary function defining whether a guest has a private or shared bathroom. For homes where a guest is in a private room, a host may be able to charge more if the room has its own bathroom.
bedrooms	[**NUMERIC**] Removed unusual properties from the bedroom count, since these are mostly tents, tiny houses, and do not fit into the bedroom model. Empty strings were then assigned a 0 value and assumed to be studio apartments
beds	[**NUMERIC**] Applied the same logic as above
price	[**NUMERIC**] Applied the substring function to remove the currency symbols so that price could be used as a number for the purposes of analysing the data. A CASE was used to add an additional logic for separating the 1000 separator. *CASE* *WHEN price LIKE '%,%' THEN CAST(replace(substr(price, 2), ',', '') AS REAL)* *ELSE CAST(substr(price, 2) AS REAL) END as price*
current_{**min/max**}_nights	[**NUMERIC**] Relabelled from {**minimum/maximum**}_nights to reiterate my own definition as the available nights for booking at the time the data was collected.
least_{**min/max**}_nights most_{**min/max**}_nights	[**NUMERIC**] Assumed from the data to be the fewest or most nights the host has rented the property at the min or max level. Changing the nights stayed in peak seasons would be interesting.
avg_{**min/max**}_nights_forecast	[**NUMERIC**] Predicted average number of nights over the next 12 months (assumed to be by month)
range_{**min/max**}_nights	[**NUMERIC**] A derived column to identify hosts that change their minimum or maximum nights stay by a large factor; hoping to identify hosts that fluctuate between long term and short term rentals
STR	[**NUMERIC**] An ordered list for listings that have short term rentals: *0 = Long term rentals* *1 = Ambiguous rentals*

	2 = Short term rentals **Note: In later views, the logic shifts to compare properties known to be short term (2) and all other rentals. Listings in the 0 and 2 categories only made up approximately 1/3 of the dataset, so including these ambiguous properties was essential for acquiring insight into short term rentals.**
availability_scale	[**NUMERIC**] Derived from the has_availability and availability_{30/60/90/365} columns, this feature defined an ordered category of low to high availability: *Null = Assumed to be inactive* *0 = No availability in the next 365 days* *1 = Availability between 90-365 days* *2 = Availability between 60 – 90 days* *3 = Availability between 30 – 60 days* *4 = Availability in the next 30 days*
availability_{30/60/90/365}	[**NUMERIC**] Number of days available in the given time period
number_of_reviews	[**NUMERIC**] Number of reviews for the given listing. Note that some properties had no date for the last_review column, but had more than one listing. This could be a way of identifying properties that have been removed and relisted (see conclusions/next steps)
number_of_reviews_{ltm/l30d}	[**NUMERIC**] The number of reviews in the last 12 months or 30 days, respectively.
recent_review_ratio	[**NUMERIC**] On the basis of the above columns , I derived a ratio to see if this could be used to identify properties that have been stayed at more recently *number_of_reviews_l30d / numer_of_reviews_ltm*
first_review, last_review	[**TEXT/DATE**] This is a fluid data type in SQLite, but must be converted to the format YYYY-MM-DD to be identified as a date. *substr(mydate, 7)* || '-' || *substr(mydate, 1,2)* || '-' || *substr(mydate, 4,2)*
duration_property_listing	[**NUMERIC**] – length the property has been on Airbnb, in years
review_scores_{}	[**NUMERIC**] Scores given by guests after staying at the property for various

	parameters. Unlikely to be used further than correlation for this particular analyisis unless a surprising interaction comes up
is_instant_book	[**NUMERIC**] Binary feature for using the instant book feature. What kind of hosts use this feature?
host_listings_count	[**NUMERIC**] Assumed to be the total number of listings across Airbnb
Calculated_host_listings_count	[**NUMERIC**] Number of properties a given host has in this dataset. Confirmed by: *SELECT COUNT(DISTINCT host_id) FROM stg_airbnb_listing*
calculated_host_listings_count_{**entire_homes /private_rooms/shared_rooms**}	[**NUMERIC**] Number of properties a given host has in this dataset of a particular room type

Other columns in the stg_airbnb_column were rremoved from the dataset. Finally, properties with no stay data were filtered out with a WHERE clause:
host_acceptance_rate is not NULL AND host_acceptance_rate != 0
AND last_review != ''

cleaned_zillow

Table column	**Working definition / logic of derived column**
zpid	[**PRIMARY KEY**] [**ID**] [**NOT NULL**] The identifier for each row in the dataset. No duplicates or nulls were permitted in this column.
city	[**TEXT**] A few properties lie outside the bounds of Austin, but may still provide some useful information about property sizes in the suburbs
street_address	[**TEXT**] The full street address is contained in some of the rows in the crime data – perhaps there is a way to join on these
street_num	[**NUMERIC**] Derived from the above column, separating the number of the street in the street address. The stg_gov_authorized_rentals dataset had "block of" information which might be a grain we can join on with this data.
street	[**TEXT**] Derived from the above column, separating the street name in the street address. This could be used as a join foreign

	key, with zip code, to access more granular crime and license information
description_length	[**NUMERIC**] The number of characters in the Zillow listing for a given property. Are there certain attributes that make a listing more compelling?
has_{**amenities**}	[**NUMERIC**] Converted into binary from True and False for the correlation matrix. Since we don't have these information about the properties in the Airbnb listings, would it be possible to find amenities that are more available in certain parts of the city
has_parking	[**NUMERIC**] A binary feature describing whether there is parking on the property. The number of spaces could not be identified with the data provided because of mismatching information

All other columns in the dataset were left unchanged for the purposes of the initial correlation matrix.

Data Analysis

View key

Below, are the views created during this project and a key of which views were used in which visualizations:

View name	Description
initial_airbnb_correlation_matrix	All data formatted as numeric, and additional logic for the short-term rental feature. (

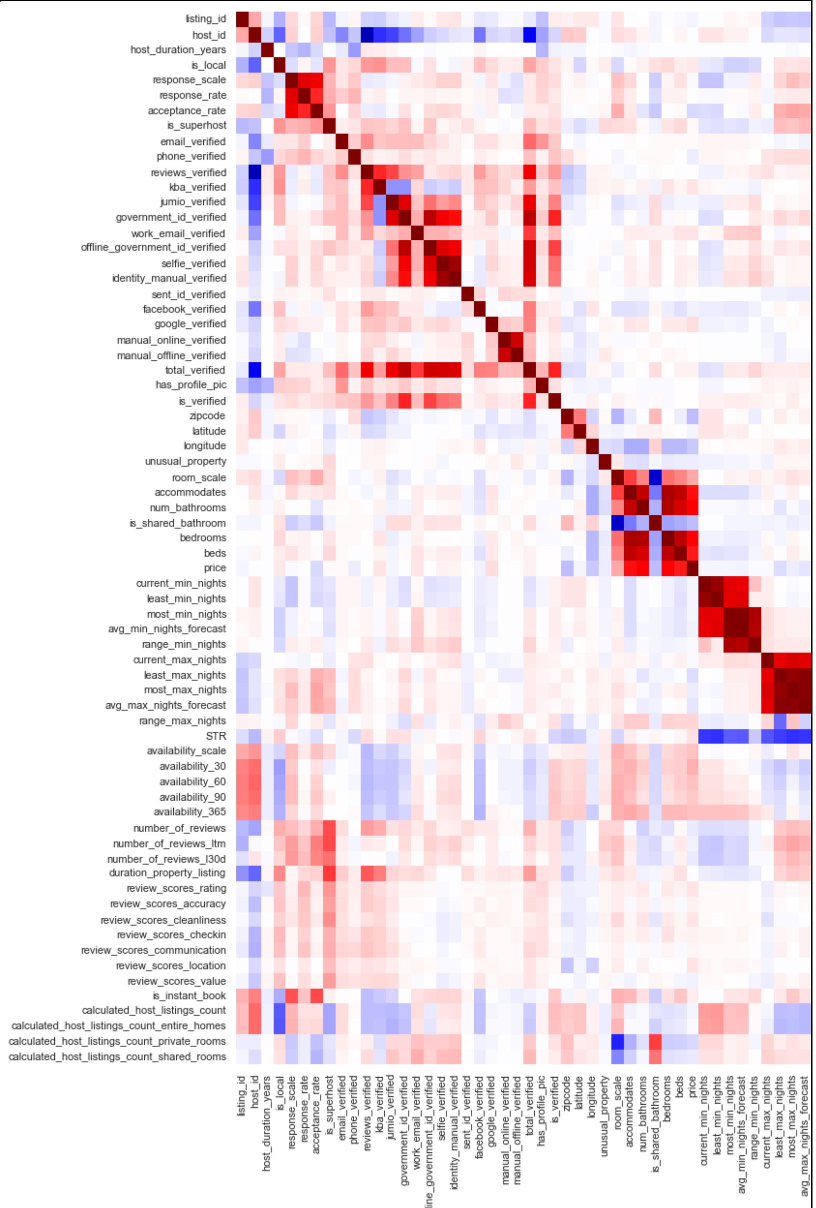
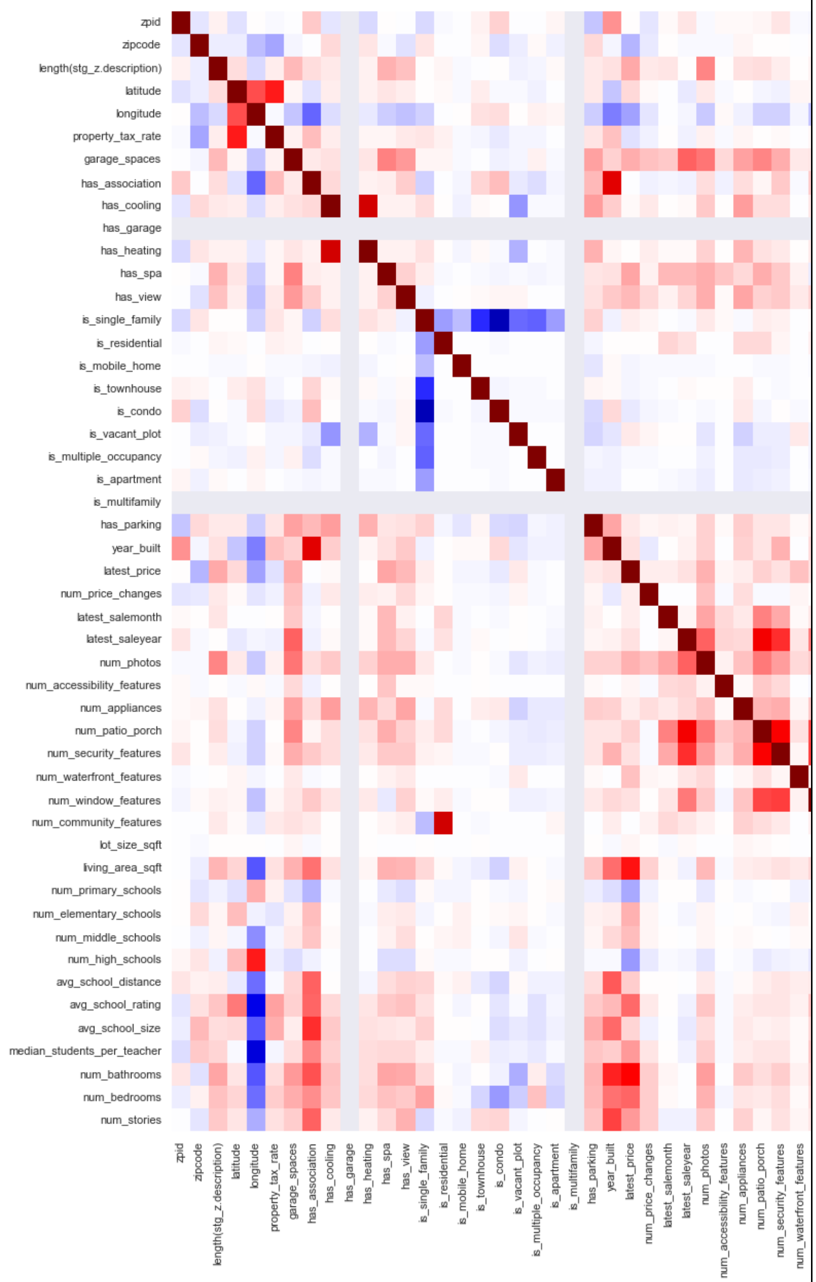

listing_id
host_id
host_duration_years
is_local
response_scale
response_rate
acceptance_rate
is_superhost
email_verified
phone_verified
reviews_verified
kba_verified
jumio_verified
government_id_verified
work_email_verified
offline_government_id_verified
selfie_verified
identity_manual_verified
sent_id_verified
facebook_verified
google_verified
manual_online_verified
manual_offline_verified
total_verified
has_profile_pic
is_verified
zipcode
latitude
longitude
unusual_property
room_scale
accommodates
num_bathrooms
is_shared_bathroom
bedrooms
beds
price
current_min_nights
least_min_nights
most_min_nights
avg_min_nights_forecast
range_min_nights
current_max_nights
least_max_nights
most_max_nights
avg_max_nights_forecast
range_max_nights
STR
availability_scale
availability_30
availability_60
availability_90
availability_365
number_of_reviews
number_of_reviews_ltm
number_of_reviews_l30d
duration_property_listing
review_scores_rating
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communication
review_scores_location
review_scores_value
is_instant_book
calculated_host_listings_count
calculated_host_listings_count_entire_homes
calculated_host_listings_count_private_rooms
calculated_host_listings_count_shared_rooms

Figure 1)

cleaned_zillow	Used as the basis of the Zillow correlation matrix ( Figure 2)
airbnb_property_segmentation_roomtype_STR	Segmentation data for Airbnb room data (Figure 3, A-D)
airbnb_union_zillow_bedroom_sqft_price	Segmentation data for Airbnb and Zillow by number of bedrooms, square feet, and price (Figure 4, A-D)
Airbnb_union_zillow_zip_aggregate_sqft_price	Segmentation by zip code and price (A-C)
zip_listings_sale_price	D
zip_78703	E
historic_zillow	F

Visualizations

Visualizations were generated with the Seaborn library of Python. See below for code snippets:

```
-- setting up libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set(rc={'figure.figsize':(20,20)})
plt.rcParams.update({'font.size': 18})


--generating correlation matrix
airbnb_host_df = pd.read_csv('airbnb_host_correlations.csv')
fig = sns.heatmap(airbnb_host_df.corr(), cmap="seismic",
annot=True, vmin=-1, vmax=1)


-- generating barplot for bedroom and listing price data,
separated into short-term and long term rentals
sns.barplot(x = 'Num bedrooms', y = 'Average listing price',
data=beds_df, hue = 'STR', palette={'LTR':'gray', 'STR':'red'})


-- generating a plot with a double x axis to compare property
listing price on Airbnb with Zillow listing price
plt.figure()
ax = sns.barplot(x='Num bedrooms', y='Most recent sale price',
data=beds_df, palette={'0': 'gray', '1':'gray', '2':'gray',
'3':'gray', '4':'gray', '5+':'red'})
width_scale = 0.45
for bar in ax.containers[0]:
    bar.set_width(bar.get_width() * width_scale)
ax2 = ax.twinx()
sns.barplot(x='Num bedrooms', y='Average sqft', data=beds_df,
ax=ax2, hatch='xx', palette={'0': 'gray', '1':'gray',
'2':'gray', '3':'gray', '4':'gray', '5+':'red'})
for bar in ax2.containers[0]:
    x = bar.get_x()
    w = bar.get_width()
    bar.set_x(x + w * (1 - width_scale))
    bar.set_width(w * width_scale)
plt.show()
```

Additional findings

Airbnb correlation matrix (continued from page 3)

5) The "is_local" column has a strong negative correlation with the calculated host listings column, suggesting that most hosts are local.
6) There is a strong positive association cluster between the host response time and response rate and accepted listings. This could indicate that the more active the host is on the platform, the more likely they are to have a successful transaction (an accepted stay); or that guests are more likely to choose hosts who are more active on the platform (directionality has not been ascertained).
7) There is also a medium positive association between the ability for guests to instant book properties and the host response time and acceptance rate. This could be in part be because contact does not need to be made with the host to accept the booking, which could explain why the response rate is not as strong an association)
8) Being a superhost is strongly associated with higher number of reviews but less strongly associated with the review rating itself. Interestingly there is no interaction with the duration of a host having a profile on the platform. Additionally, a superhost is correlated with having fewer properties (identified by the calculated host listings column, which identifies Austin listings).
9) There is a strong positive association between the number of reviews and the duration of the property listing on Airbnb.
10) The strongest association between host verification methods appear to be the government_id (both online and offline verifications), and most hosts use this method to verify their identities. However, this does not seem to have any association with price.
11) Room scale, which increases in value from shared rooms to entire properties, is strongly associated with the accommodates column, which identifies the total number of guests able to stay in a property. This finding suggests that more people are able to stay in entire properties than in shared rooms (and that larger groups may prefer to stay in larger properties).

Zillow correlation matrix (continued from page 23)

4) Schools with higher ratings are larger, and have more students per teacher than schools with lower ratings. What is driving this effect? It would seem more likely that smaller schools with a higher access to teachers would have better ratings. Is this just a feature of there being more students in the school?
5) A strong association exists between location (latitude) and property tax rates, suggesting the northern part of the city has a higher tax rate, which could act as a deterrent for investors.
6) More recent property sales have more photos in the listing, more security features, and more windows. Are buyers wanting more security features? Has crime been increasing in Austin? There might be a time component here that would be worth looking into further.
7) Schools with higher ratings are larger, and have more students per teacher than schools with lower ratings. What is driving this effect? It would seem more likely that smaller schools with a higher access to teachers would have better ratings. Is this just a feature of there being more students in the school?