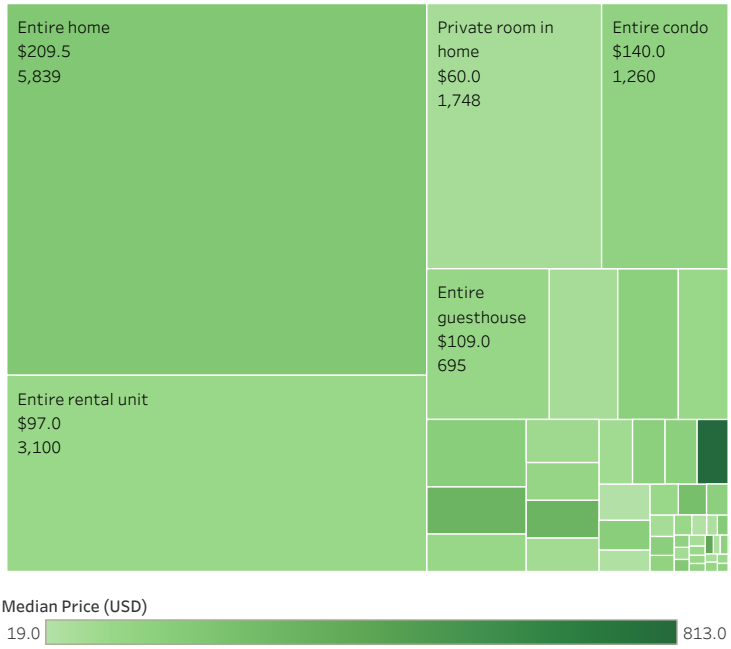
AUSTIN, TX / AIRBNB MARKET ANALYSIS

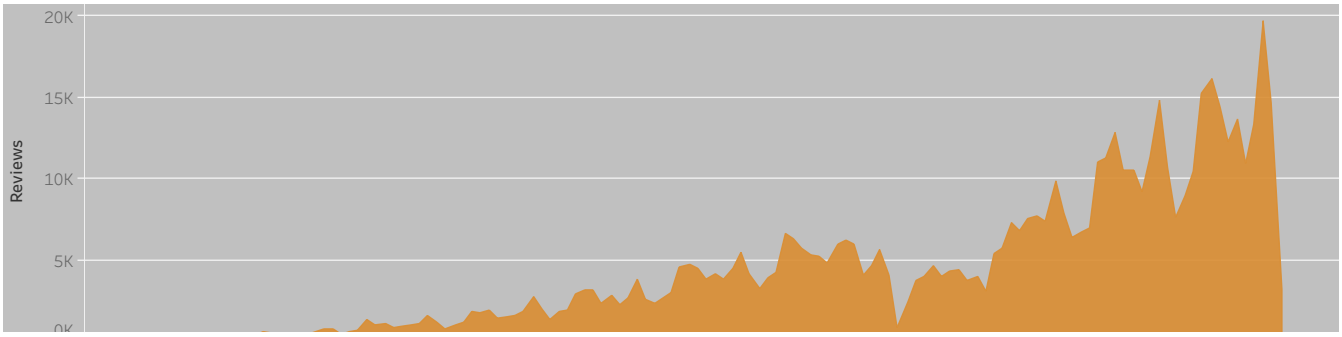
As Austin's reputation as a tech and tourism hotspot continues to soar, another trend has gained momentum: **the rise of Airbnb and short-term rentals**. With visitors flocking to experience the city's unique culture and events, the demand for alternative accommodations has surged. Airbnb listings offer travelers a diverse array of options.

As Austin embraces innovation in all its forms, the proliferation of Airbnb and short-term rentals further solidifies its status as a dynamic destination for both business and leisure travelers alike.

With a webscrape of publicly available data from Airbnb, we can look deeper into the current state of the ATX short term rental (STR) market.

Property Type Counts with Median Price per Night (Dec '23)





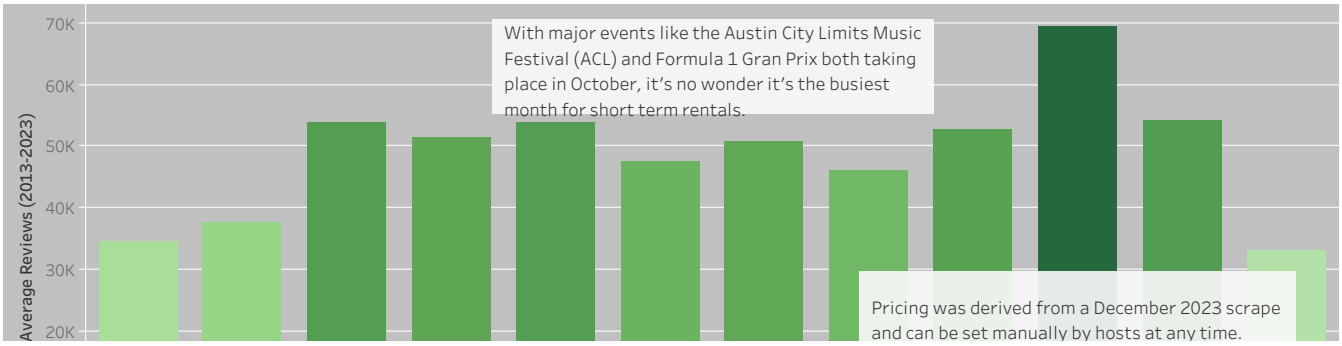
Short Term Rentals - Not Just Airbnb

While Airbnb commands a substantial share of the online travel agency (OTA) market, it operates within a competitive landscape alongside platforms such as Vrbo, Booking.com, and Expedia. Despite this competition, Airbnb stands out as the leading player in the hotel-alternative sector.

Short term rentals (STR's), also called vacation rentals, are defined as rentals with durations less than 30 days.

The methodology used to estimate revenue acknowledges but cannot fully account for bookings thru other services. This, plus the web scrape occuring in the winter/offseason, means that estimated revenues are likely conservative, especially compared to the peak months of March- October, or if a given listing does a lot of business thru Vrbo.

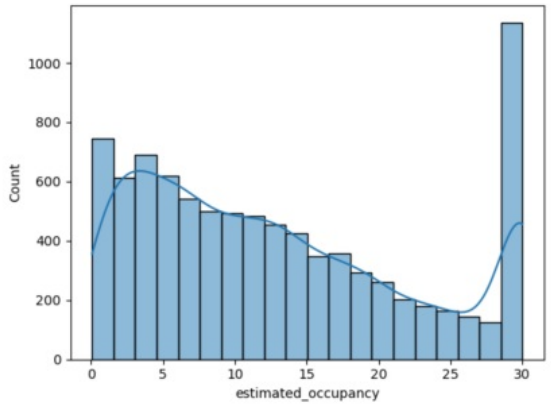
ATX STR Seasonality



About the Data

Downloaded from Inside Airbnb, (not endorsed by Airbnb or its competitors) an data/advocacy site with the mission to measure the impact that Airbnb has on affordability and housing in major tourist destinations across the globe.

Inside Airbnb conducts quarterly webscrapes of publicly available data from Airbnb, and therefore requires interpretation for key figures such as occupancy and revenue, as this information would be limited to hosts and internal Airbnb metrics.



Above: Seaborn histogram of entire_home listings by estimated occupancy, derived from each listing's reviews per month and minimum stay length.

2. Data Wrangling

- Since we don't have an outright revenue column, we can derive it borrowing from the methodology outlined in Inside Airbnb's assumptions page: http://insideairbnb.com/data-assumptions/.

- We'll derive an ['estimated_occupancy'] metric by dividing ['reviews_per_month'] by 50% (in other words x2), times the minimum length of stay if greater than 3 nights, and capped at 3 if minimum stay lower than 3 nights.
    - not every guest will review. Airbnb CEO Brian Chesky uses 72% metric, which is likely optomistic; New York attourney general uses review rate of 30.5%. Inside Airbnb opted to split the difference with a 50% reviews to estimated bookings rate.
    - listings.csv dataset offers several 'minimum nights' metrics, including:
        - 'minimum nights', minimum night stay for the listing
        - 'minimum_minimum_nights', the smallest minimum_night value from the calender (looking 365 nights in the future)
        - 'maximum_minimum_nights' the largest minimum_night value from the calender (looking 365 nights in the future),
        - minimum_nights_avg_ntm, the average minimum_night value from the calender (looking 365 nights in the future)
    - opting to use 'minimum_nights_avg_ntm' for occupancy calculations. It may be slightly more accurate than 'minimum nights'
    - ['estimated_occupancy'] capped at 30 days for full occupancy.
        - Inside Airbnb caps occupancy at 70% to remain conservative; their goal is to demonstrate the effect which Airbnb's might impact the local housing supply.

- Next we'll derive ['estimated_revenue'] by multiplying ['price'] (*which can be changed by the host on a given date but for our purposes will remain static*) times ['estimated_occupancy'].

From Juypter Notebook 6.3, an explainer on the methodology to derive estimated occupancy and estimated revenue.
By no means a perfect estimation, this derivation can still give us an idea on what listings are overperforming in the market.

See github link here:

Austin, Texas, United States

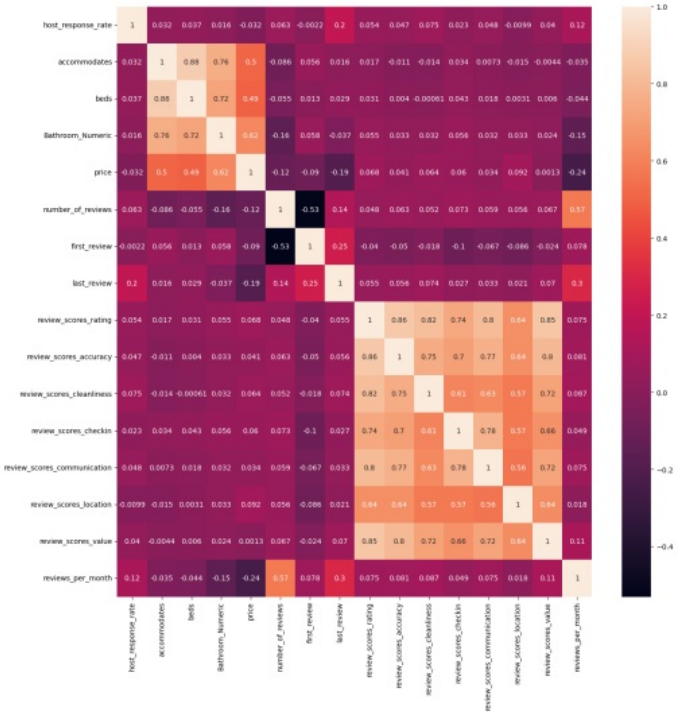
15 December, 2023 (Explore)

Country/City	File Name	Description
Austin	listings.csv.gz	Detailed Listings data
Austin	calendar.csv.gz	Detailed Calendar Data
Austin	reviews.csv.gz	Detailed Review Data
Austin	listings.csv	Summary information and metrics for listings in Austin (good for visualisations).

Worked primarily with 'listings.csv.gz' dataset, but also utilized the 'neighbourhoods.geojson' for geospatial analysis using folium python library, as well as the reviews.csv for time series data.

Pre-data cleaning, the Dec 23 detailed listings dataset

Exploratory Data Analysis



Correlation heatmap did not yield very strong or unexpected correlations.

Besides the diagonal of 1's that correspond to a variable lining up with itself, we see a couple of positive correlation zones:
- 'beds' and 'accomodates', which makes sense.
- The 7x7 square of 'review_scores' and its subcategories; a high score in one would generally be correlated with other high scores.
-Number of reviews and reviews per month. Not a very strong correlation but stronger than the surrounding, enough to take notice.

In terms of negative correlations, bearing in mind that we don't see any highly negative correlations in this heatmap:
- Price and host response rate; so the higher the price, the lower the response rate. I could buy that; more expensive properties, probably more vacancy/ less stays, balanced out by more money per stay. Maybe these listing owners only rent it out occasionally, on certain event weekends where they can command higher prices.
-First review and number of reviews; So the older(lower number) the first review, the higher the number of reviews in total.

```
df.loc[df['price'] < 100, 'Price Category'] = 'Low price'
```

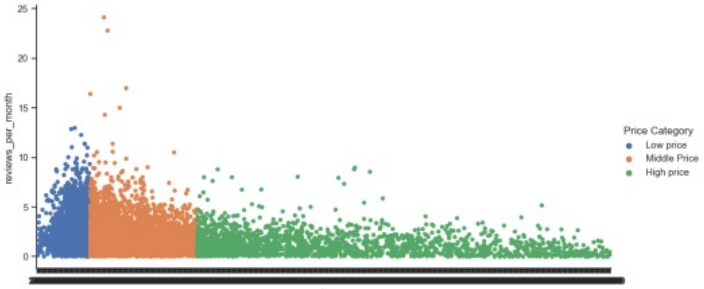
```
/var/folders/12/fz0w81fd1mb1hklwgzyg3bx00000gn/T/ipykernel_79224/2597053395.py:1: FutureWa
ning: Setting an item of incompatible dtype is deprecated and will raise in a future error
of pandas. Value 'Low price' has dtype incompatible with float64, please explicitly cast t
a compatible dtype first.
  df.loc[df['price'] < 100, 'Price Category'] = 'Low price'
```

```
df.loc[(df['price'] >= 100) & (df['price'] < 250), 'Price Category'] = 'Middle Price'
```

```
df.loc[df['price'] >= 250, 'Price Category'] = 'High price'
```

```
df['Price Category'].value_counts(dropna = False)
```

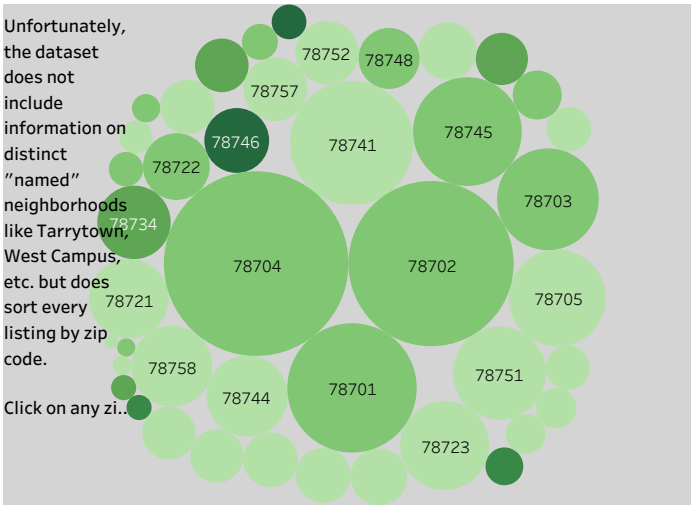
```
Price Category
Middle Price    4337
Low price       2187
High price      2150
Name: count, dtype: int64
```

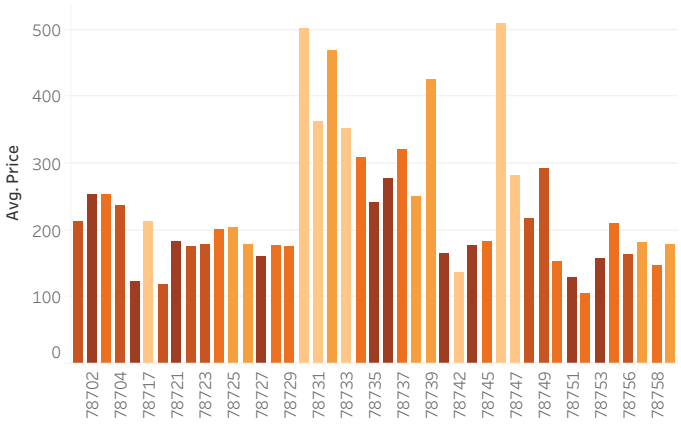


Austin Texas by Zip Code

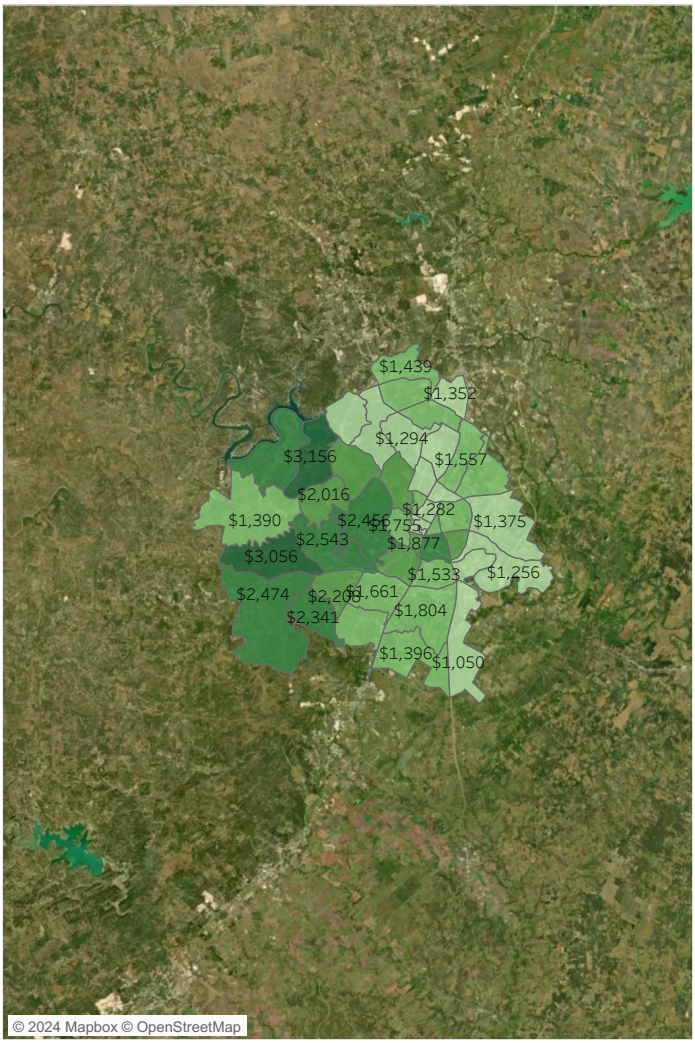
Unfortunately, the dataset does not include information on distinct "named" neighborhoods like Tarrytown, West Campus, etc. but does sort every listing by zip code.

Click on any zi..



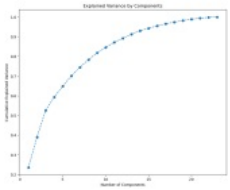


Avg. Estim.. 6.037 12.892



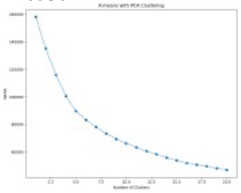
Clustering: Unsupervised Machine Learning Technique

Using **Scikit-learn**, a machine learning library for Python, I analyzed the listings dataset and determined that 80% of its variance came down to 9 components.

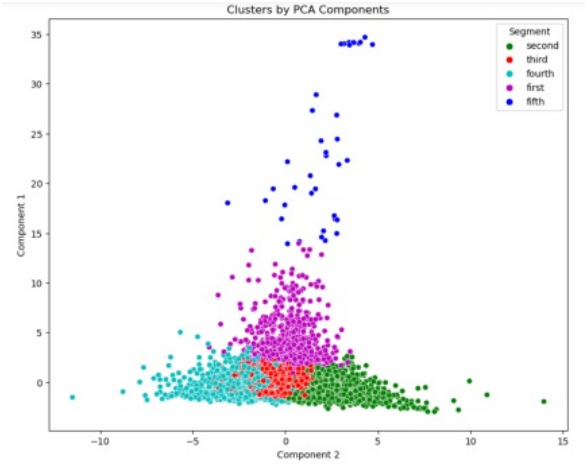


From there, I standardized the data using StandardScaler(), as the different components (pricing, review scores, number of reviews, age of listing in months) all had different orders of magnitude.

Then I applied PCA (Principal Component Analysis) followed by a k-means clustering algorithm in order to reduce the datasets dimensions and determined the number of clusters, which was 5 segments, via the elbow test.



After plotting a number of variable pairs with their clusters, and aggregating descriptive statistics for key variables to distinguish the segments, I was able to derive some key takeaways that distinguish the segments from one another. ->




```
1 df_pca_kmeans['Segment'].value_counts()

Segment
third     4357
second    1989
fourth    1287
first      733
fifth       38
Name: count, dtype: int64
```

Takeaways from Descriptive Statistics:

first segment - new and mediocre listings

- 2nd lowest listing age statistics (in months), only ahead of fifth segment.
- 2nd lowest median estimated revenue, 2nd lowest review scores.

second segment - established, low margin, high volume.

- highest mean number of reviews by a fair margin, and yet only 3rd in review numbers by median (3rd and fifth categories have wide margin between median and mean)
- lowest average prices by mean, and 2nd lowest (to first segment) in median.
- High review scores, but still 3rd behind fourth and third segment.
- Second highest estimated revenue, behind fourth segment, in both mean and median.

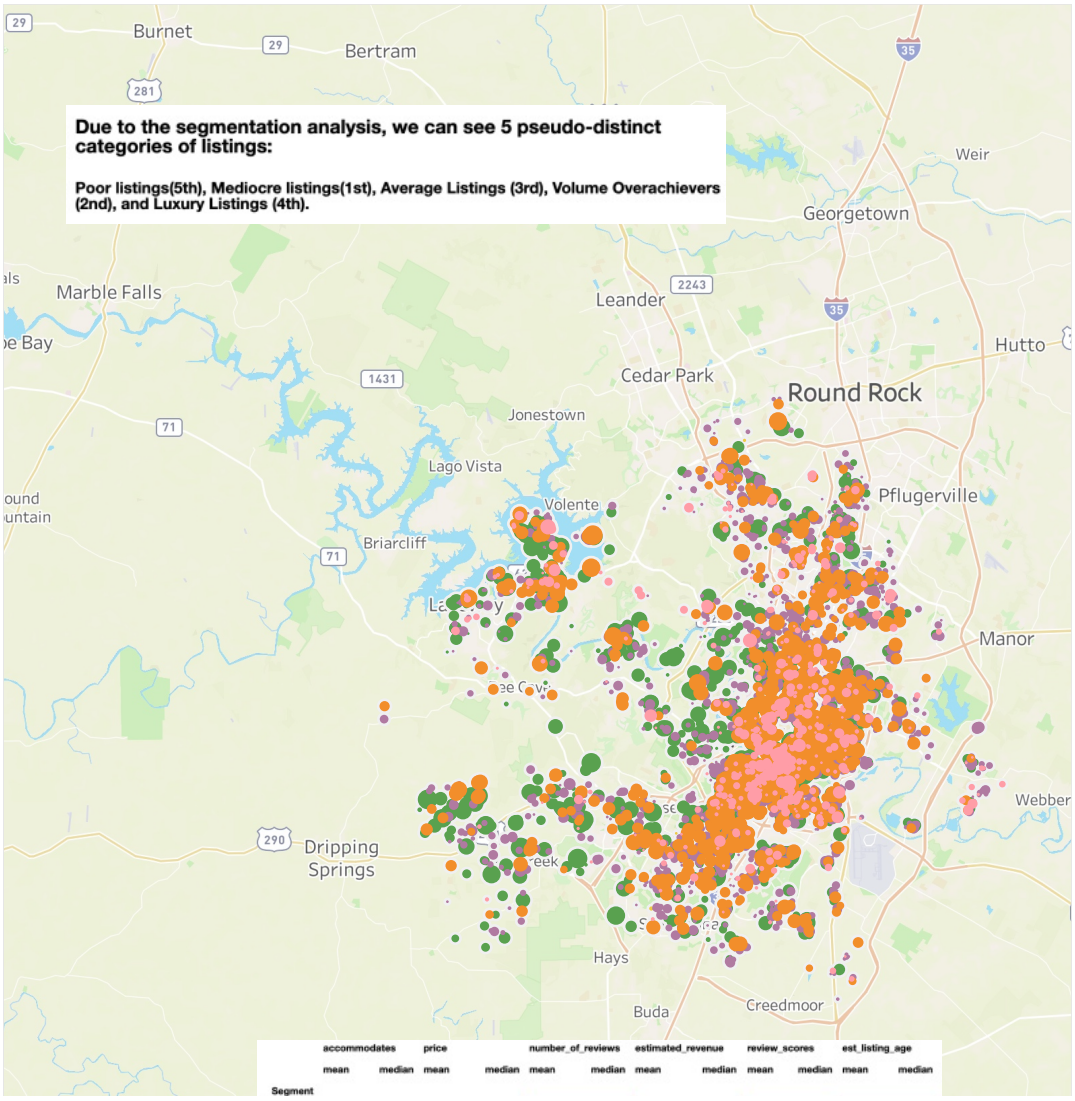
third segment - middle of everything, checks the box.

- middle in price by mean
- middle by review count
- middle estimated revenue
- curiously, highest review scores
- Also tied for highest median listing age, close first for mean listing age.

fourth segment - luxury/upscale/ large listings

- highest accomodation numbers by a wide margin: 10+ vs 4-5 for the other segments
- Highest price by a wide margin: median 407 vs other segments in the 100s.

	accommodates		price		number_of_reviews		estimated_revenue		review_scores		est_listing_age	
	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median



Segment
- first
- second
- third
- fourth
- fifth

Estimated Revenue
- 100
- 2,000
- 4,000
- 6,000
- 8,000
- 10,000

1st segment listings, mediocre but also newer (less months since first review), appear most concentrated in the city core.

The central core of the city is dominated by 2nd segment listings, the volume overachievers. These overachieving listings appear all over the city, a role that I expected only the 3rd segment, or average listings, to hold.

3rd segment, or average listings, as one might expect are across the city, but they seem overcrowded by volume overachievers across the board.

Fourth, or luxury listings are difficult to isolate without phasing out the other segments, but appear in the main city as to the west of the city, especially West Lake Hills. These listings too, seem positioned more diversely than expected.

The 5th segment, poorly reviewed underperformers, are non-factors.

Conclusion and Recommendations

Across Austin, there is a robust inventory of high performing listings, average listings, and only a handful of not so high performing listings.

On a zip code/ neighborhood level, there are trends:

- 78704 and 78702 stand out as the most dense by listing count, and could reasonably be deemed the hub of Airbnb activity in the city.
- Neighborhoods to the west of this hub, especially 78746, 78732, and 78734, showed some of the highest prices per night and revenue figures on average.

On a city-wide level:

- we can see a mass of listings in the central city and a slightly lesser but still significant STR presence to the North, South, West and Southwest of Austin proper.
- relative lack of development to the East of Austin is noteable.

From our machine-learning cluster analysis, we determined two distinct routes toward listing success:

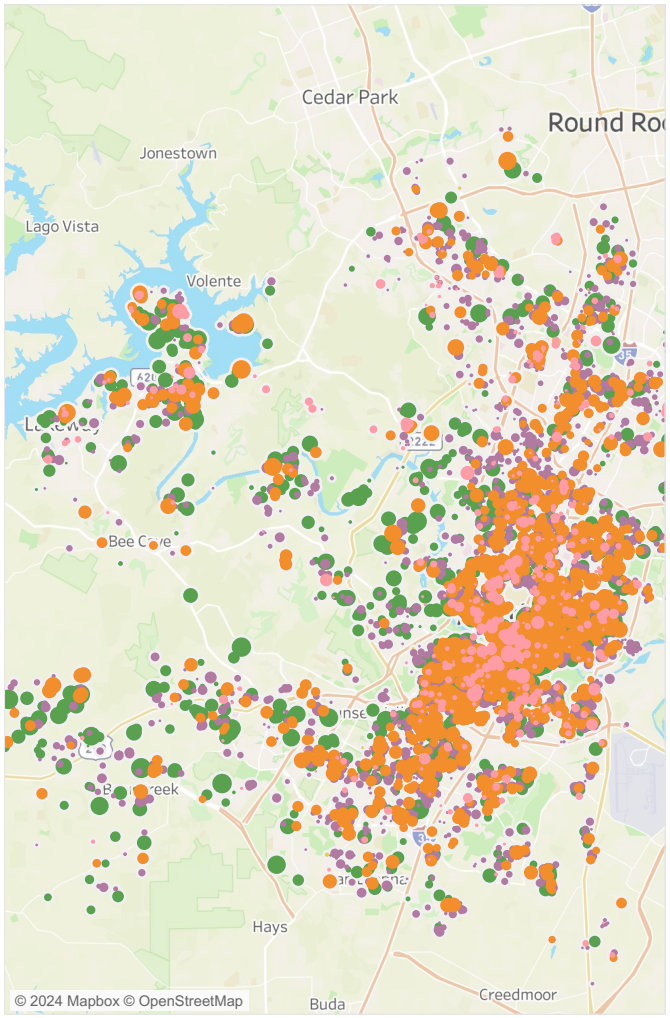
-Volume bookings at a competitive price or large and luxury listings.
- Surprisingly, we saw listings of these segment types coexisting in almost every neighborhood.

Limitations of this market analysis:

- lack of actual revenue figures in the data set necesitated an estimation based on review rates, which is not exactly representative of actual bookings.
- some places have better review rates than others, which can be a function of host involvement/prompting, listing category, or just chance.
- the 'bedrooms' variable did not come across in this scrape, so 'bathrooms' and 'accomodates' figures had to be used as proxies for listing size.
- the figures cited are based on an 'offseason' scrape, likely all prices and therefore revenue would be higher in a scrape during peak months.

Recommendation:

- for any would-be STR investors, I would caution taking on an investment in the ATX area as it appears to be saturated.
- plenty of demand over the years has created a vast supply. Listings must be pareto-efficient to stand out.
- could still be a serious appreciation play, but with high prices and mortgage rates, cash flow will be harder to come by.



Estimated Revenue Segment