Data Storytelling Project

With Capstone Dataset

By Joshua Karpen

This report details the data visualizations I created while analyzing my capstone dataset. As a reminder, I am studying AirBnB data, combined with Yelp data, and attempting to predict price and utilization. The full code used to create these visuals is available in a Jupyter Notebook on my GitHub page here: https://github.com/jkarpen/Springboard_Projects/blob/master/Capstone/Scripts/Capstone_Data_Story_Script.ipynb

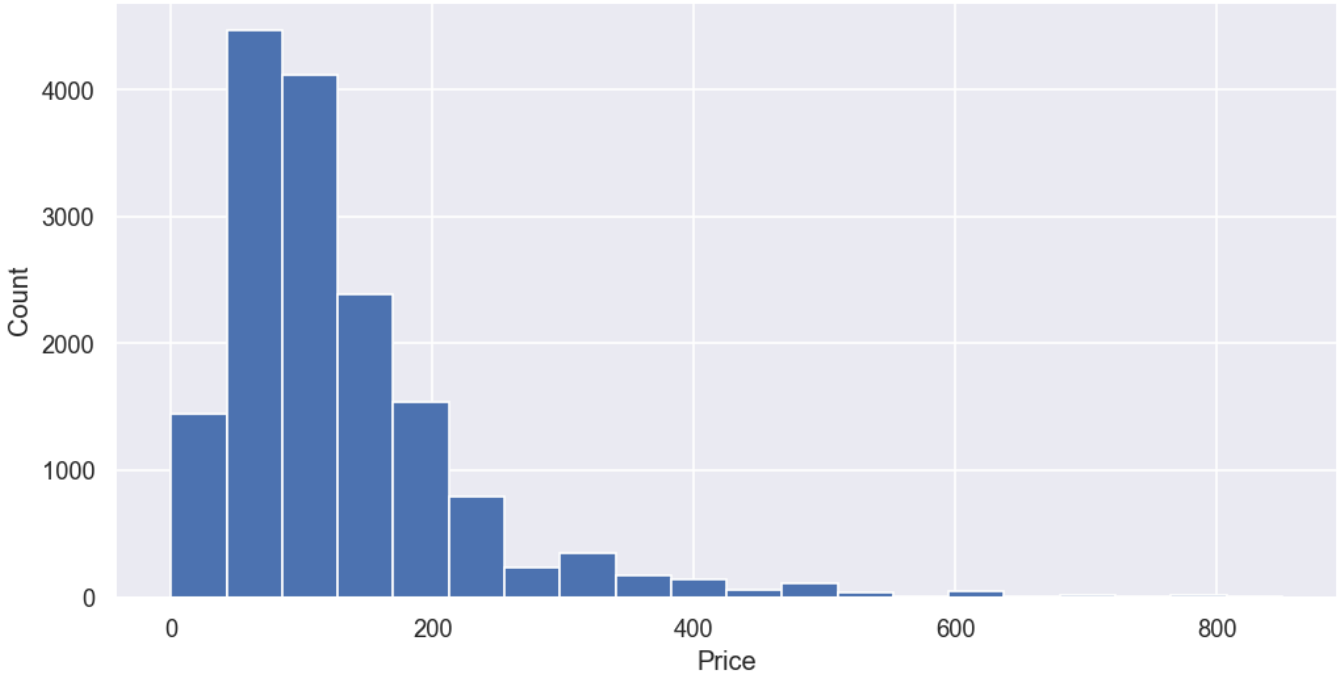
Questions Asked

With close to 60 variables in the dataset after cleaning and merging with Yelp data, I knew I had to focus my exploratory data analysis by coming up with a set of questions before diving into the data. I came up with the following questions:

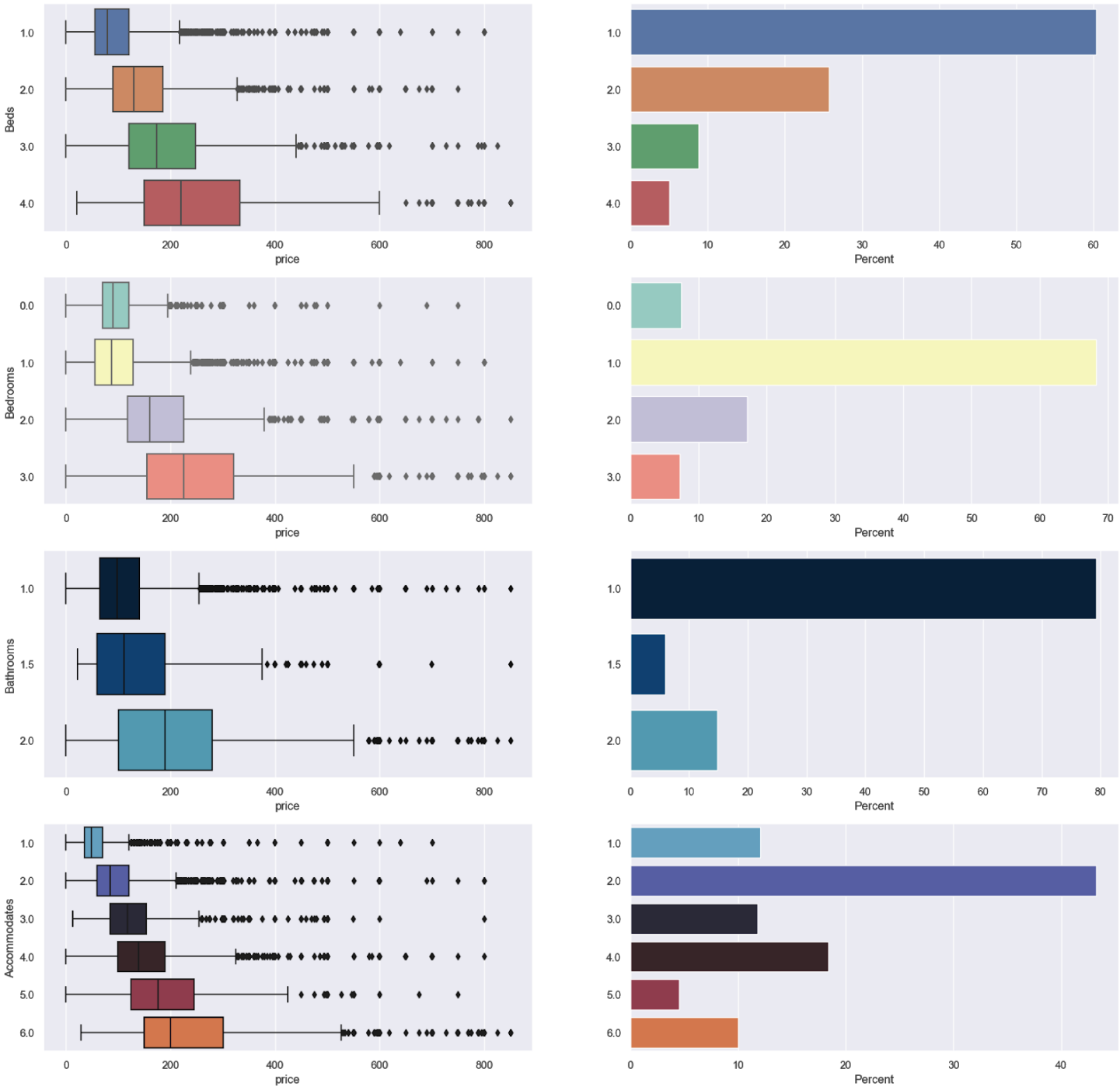
- What is the most common host response time?
- What is the distribution of host response rate?
- How many superhosts are in the dataset?
- How many properties are listed with their exact location (using is_location_exact)?
- What are the most common property, room, and bed types, and combinations of each?
- What is the distribution of accomodates, bathrooms, bedrooms, beds?
- What is the distribution of the cost variables, other than price (security_deposit, cleaning_fee, and extra_people)? And how do they relate to each other?
- What is the distribution of minimum_nights?
- How often are people updating their calendars (based on calendar_updated)?
- What is the distribution of the four availability metrics, and how do they relate to has_availability?
- How many properties have turned on the instant bookable feature?
- What are the most common cancellation policies?
- How often do listings require guests to provide a profile picture and/or phone verification, or both?
- What is the distribution of reviews per month? How does it relate to the availability metrics?
- What are the distributions of the Yelp metrics (business count, total reviews, and avg. rating), and how do they relate to each other?
- How do all of the above variables relate to price, my target variable?
- How is price distributed by geographic location? This would require plotting on a map. If possible it would be nice to create something interactive where I could view some of the other variables by location as well.

I created visuals to address all of these questions. I decided not to include the basic distribution and category count charts unless they had something interesting to show, but all charts can be viewed in the Jupyter Notebook linked above. I have divided the visuals into categories. Facts About Listing covers data points related to the listing itself, such as number of bedrooms, bathrooms, how many people are allowed to stay.

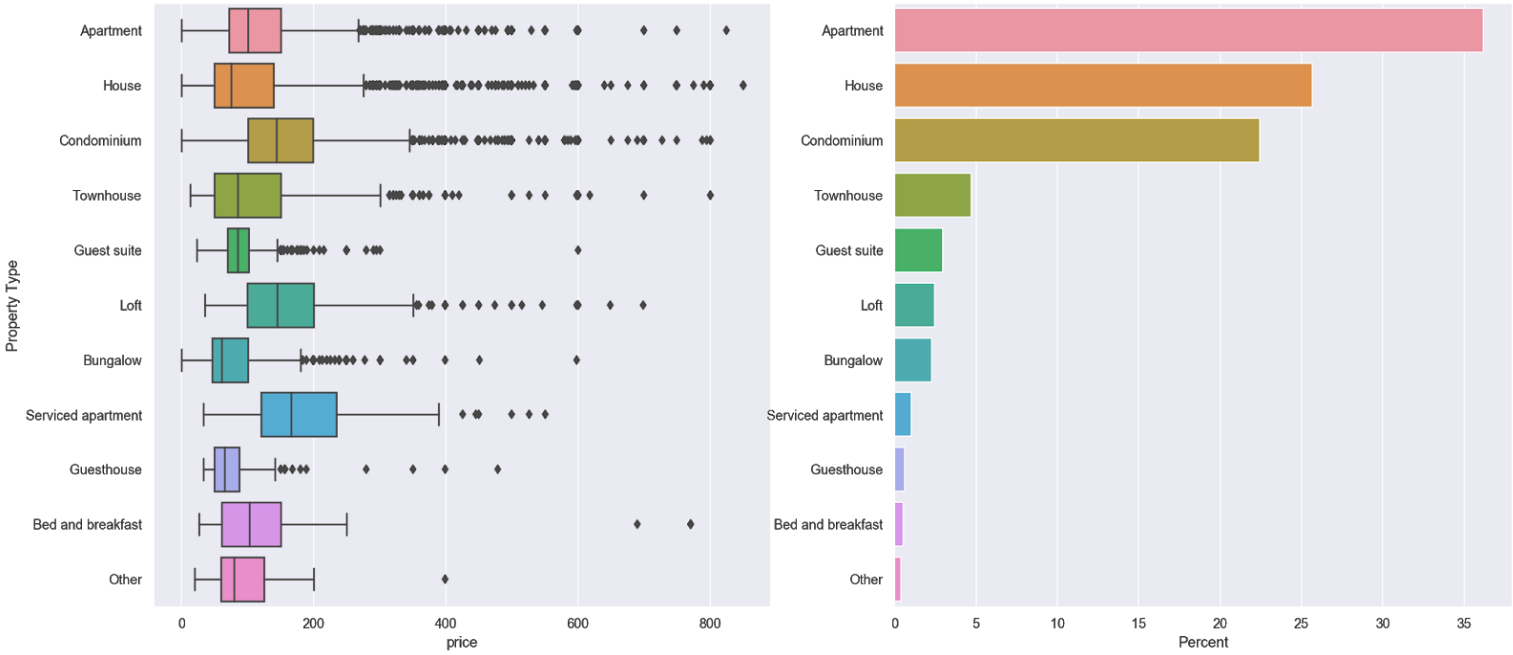
Facts About Listing



- Figure 1 - Distribution of Price
    - Most listings have fall below $200/night
    - The peak appears between $50-150 range.

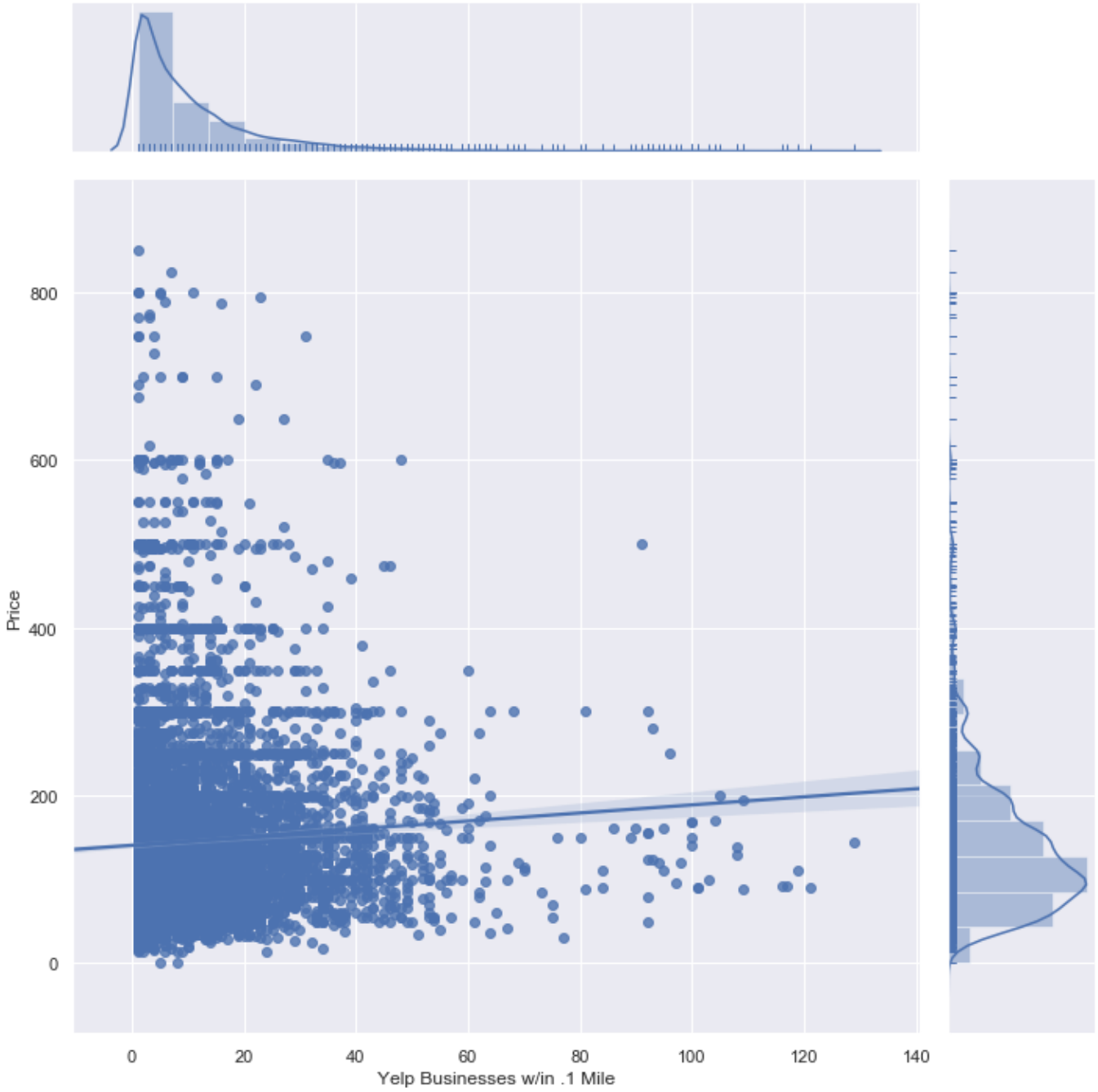


- Figure 2 - Beds, Bedrooms, Bathrooms, and Accommodates - Box plot against price (right side) and proportions (right side)
    - This chart shows that as the number of beds, bedrooms, bathrooms, and number of people accommodated increases, the average price increases as well (demonstrated by the box-plots on the right).
    - 2 people is the most common number of accommodations by far, followed by 4 people.

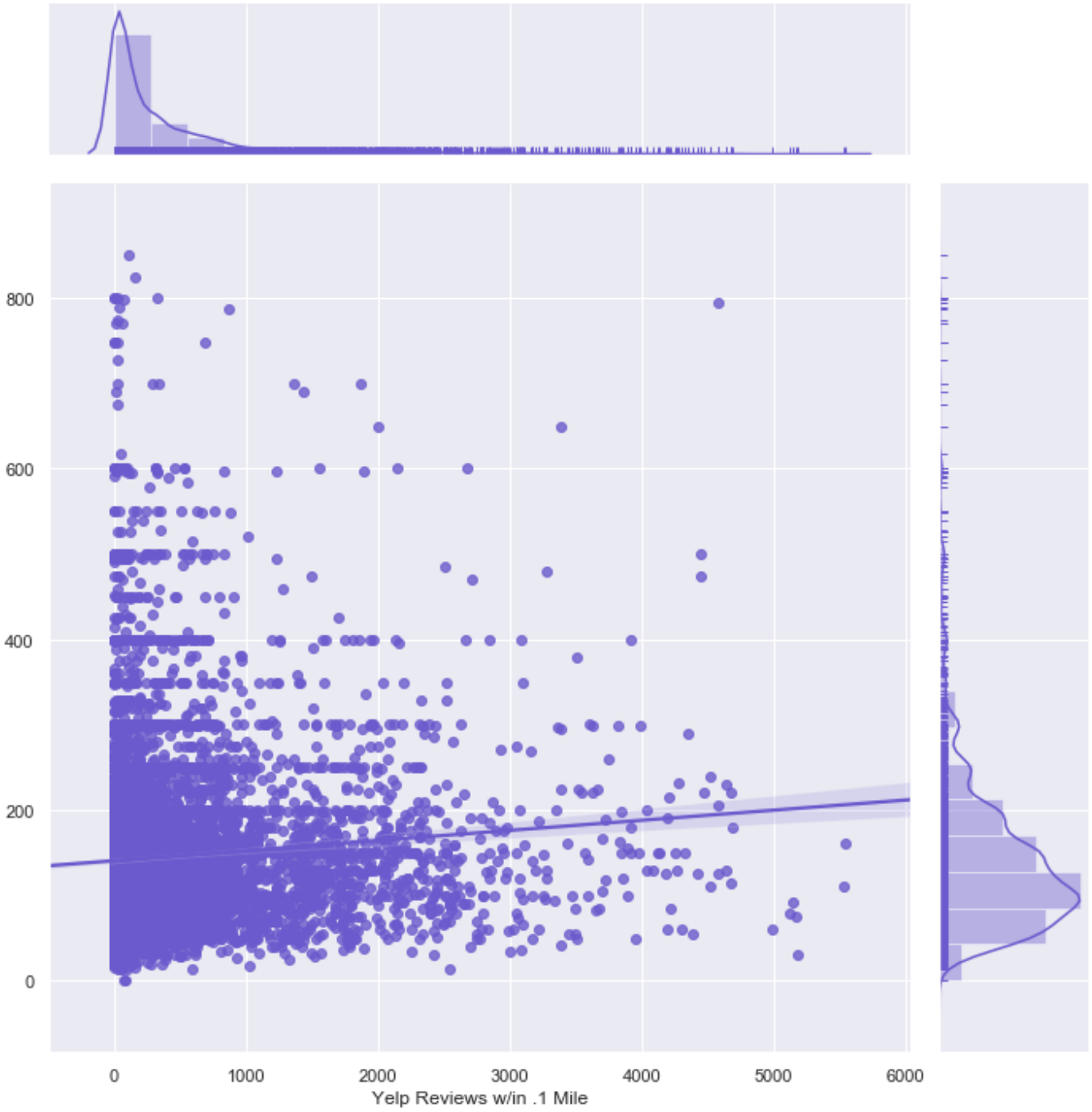


- Figure 3 - Property Type - Box plot against price (right side) and proportions (right side)
    - By far the most popular property types are apartments, homes, and condos. Townhouses, guest suites, lofts, and bungalows make up a much lower proportion but are still somewhat significant. Then there are ten or so property types that have only a handful of listings each, with “interesting” names like Cave, Aparthotel, Castle, Treehouse, Earth house.
    - Interestingly, Condos have a higher average price than apartments or homes. This might be due to where they are located.
    - The highest average prices are charged by boutique hotels, but the locations with the highest individual listing prices are still apartments, homes, and condos.

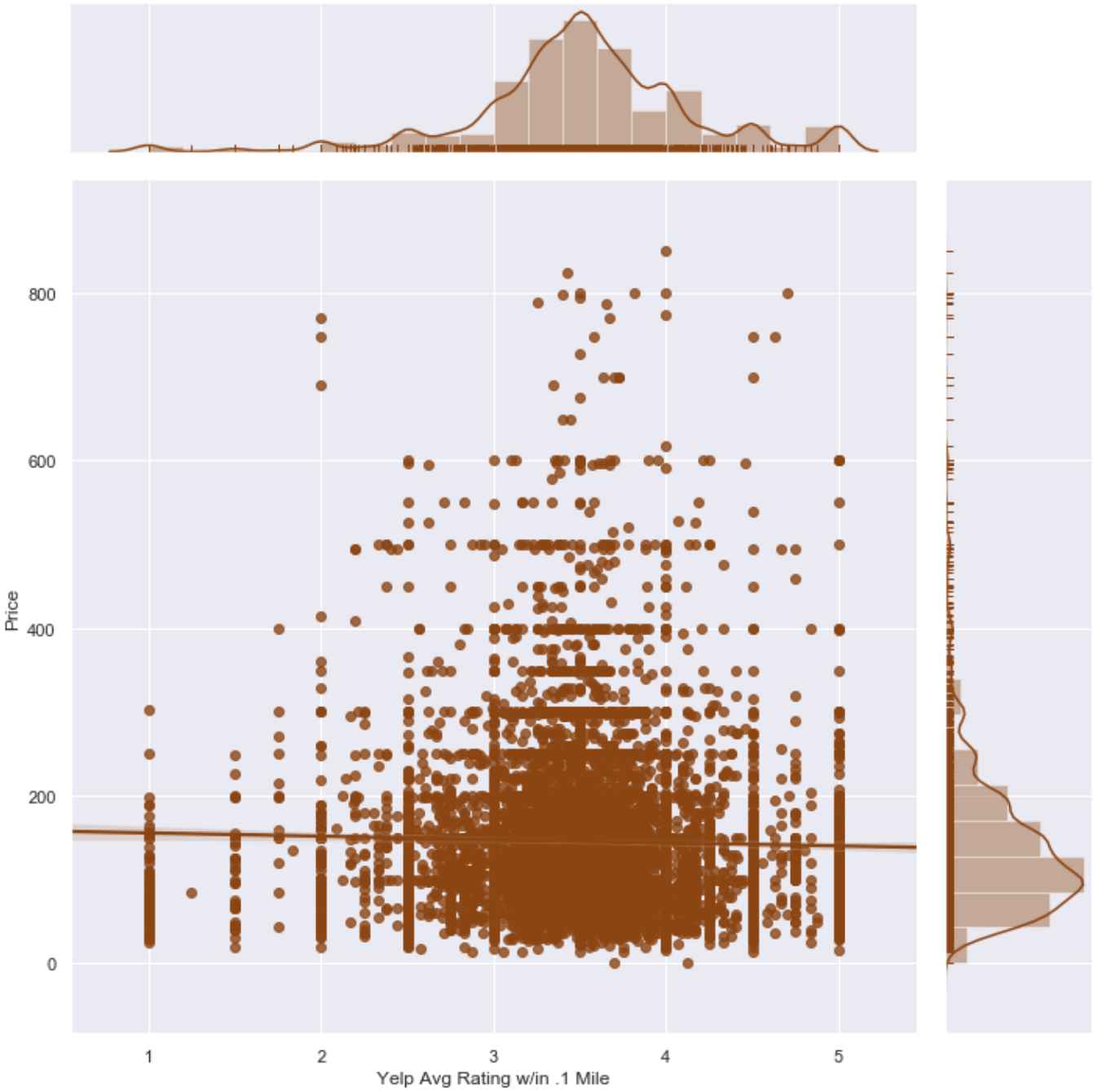
Yelp Business Data



- Figure 4 - Count of Yelp businesses against listing Price
    - As the number of businesses within .1 mile increases, the price increases as well
    - The majority of listings have fewer than 20 businesses within .1 mile

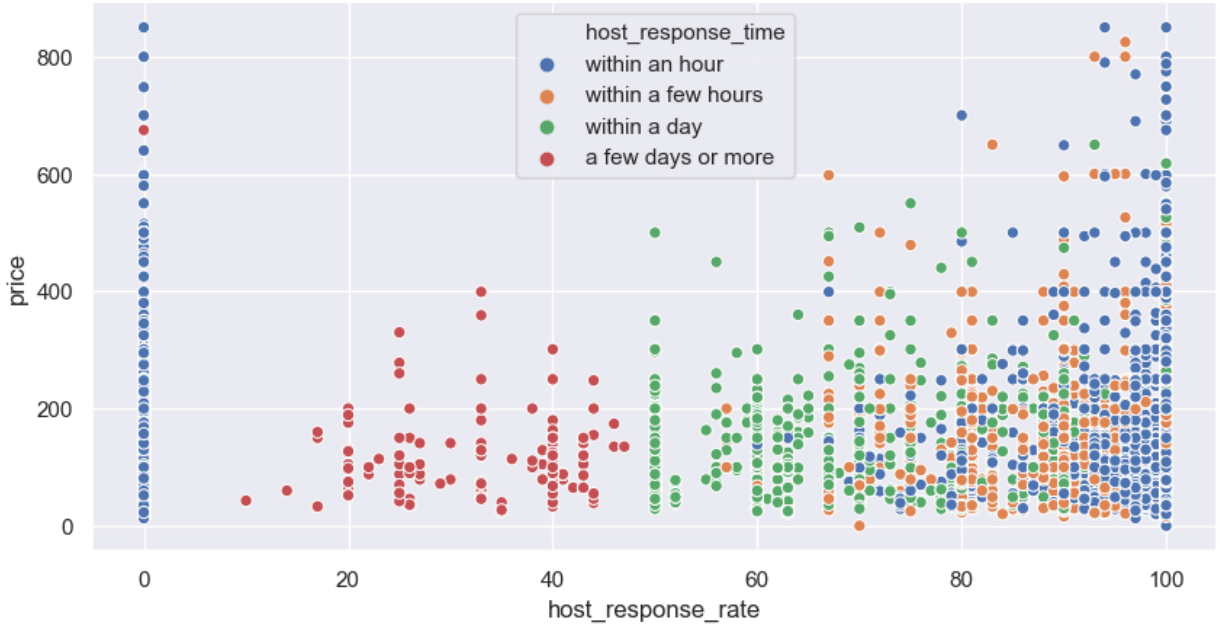


- Figure 5 - Total Reviews of Yelp businesses against listing Price
    - Listings near businesses that are have a high rating count are able to charge more, though the trend is weaker than that shown in the business count plot
    - Most businesses have fewer than 500 reviews

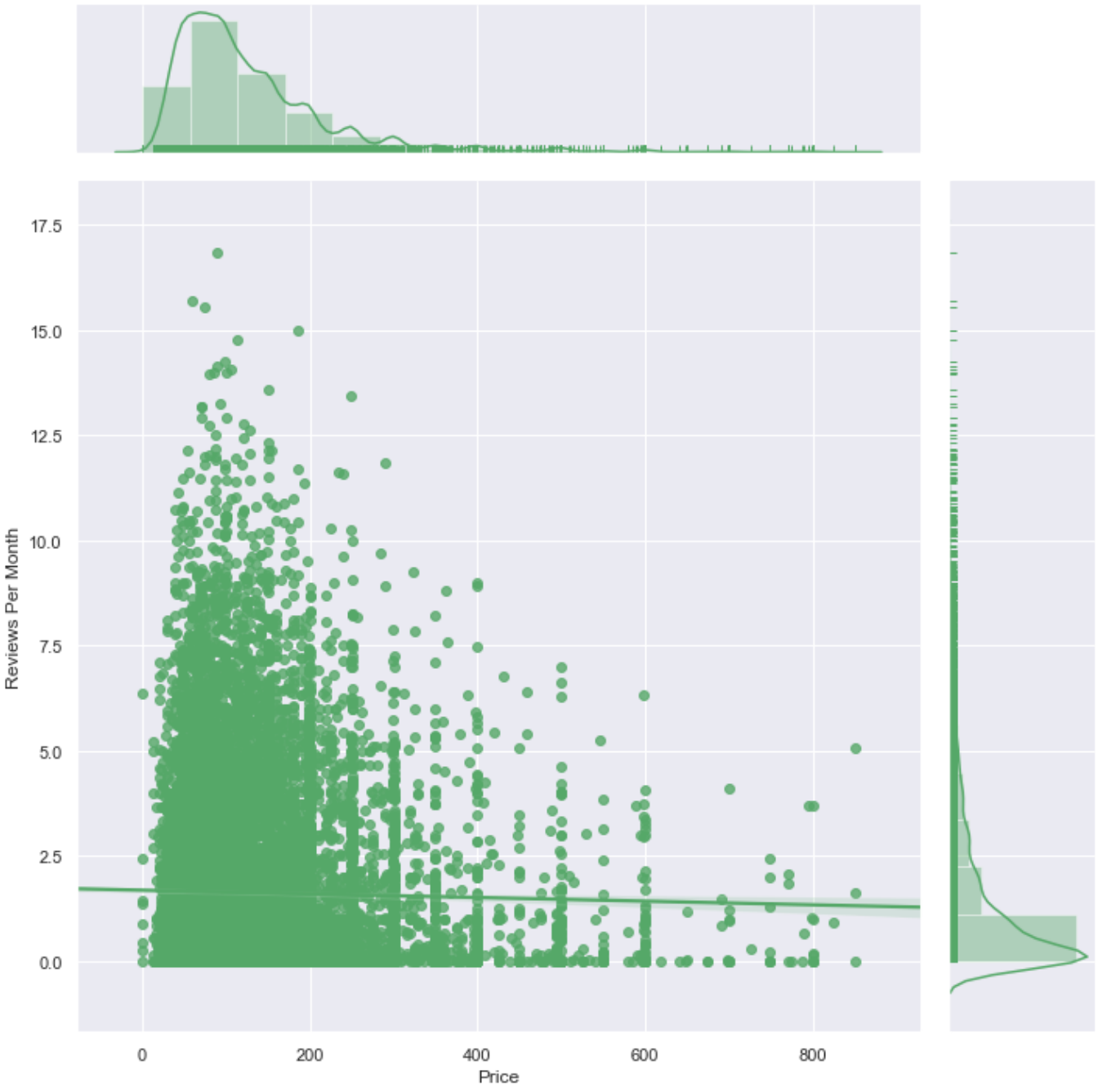


- Figure 6 - Avg. Rating of Yelp businesses against listing Price
    - Average rating appears to have the weakest relationship of the Yelp variables, based on the trend line.
    - The trend may not be best described by a linear model. Instead something curved may be more appropriate.

Interactions with Host



- Figure 7 - Host Response Rate vs. Price, with Host Response Time on color
    - Hosts with higher response rate also tend to respond more quickly, and the better a host is at responding, the more they charge.
    - They might charge more because they are better at responding to potential customers, but more likely they respond more often and quickly because they know they can make more money than someone charging a lower price.



- Figure 8 - Price (x-axis) against Reviews Per Month (y-axis)
    - As the price increases, the number of reviews_per_month, which is an estimator for how heavily booked a listing is, appears to decrease.
    - The relationship is weak, as seen by the nearly flat trend line.