The data for this project comes form Kaggle. It is a dataset of about 74,000 observatiosn and 29 variables. However, after cleaning the data to rid it of outliers and NA values, it is down to about 47,000 observations and 23 variables. I will work to clean this better as I try to make my model better in the coming weeks.

1. Can you count something interesting?

Because this is a dataset of Air Bnb units in the united states, I decided to see in general where each is located by counting the number of observations in the ‘longitude’ column that are less than or equa to ~-90. This is a lose approximation for the longitude of the Mississippi River, and there are 19,126 (40.4%) units that satisfy the criteria. This means that about 60% of the units are on the east side the River.

1. Can you find some trends?

Keeping with an analysis of location, I set out to see if there was a relationship between the price of a unit and its longitude. I counted how many units were priced below the average ‘log\_price’ (~4.751) AND west of the Mississippi. I did the same for the east side of the river. The results were 9,697 and 14,993 units, respectively. Put into percentage terms, about 50.1% of the units west of the Mississippi were priced below the mean and about 53.3% of the units east of the Mississippi were below the mean. Not a lot to report. However, if counting how many total units were below the mean, I realized that more than half of all units were below it, meaning that the mean is lower than the median. Thus there are either a few high or a few low prices that could be outliers.

1. Can you make bar plot or a histogram?

To continue with the above-mentioned analysis, I decided to create a histogram of ‘log\_price’ to get a feel for the distribution of the variable. Given that the mean was below the median, but close to it, I hypothesized that it would be a bell-shape with a slight skew to the right. I also made a second histogram to layer the distribution of ‘log\_price’ and ‘review\_scores\_rating’. The latter is not normally distributed, which presents an interesting conflict of price and quality.

1. Can you compare two quantities?
2. Can you make a scatter plot?

I made two scatter plots. The first one measures longitude on the x-axis and latitude on the y-axis. Its purpose is the show clusters of where each unit is located. The result shows that only 6 areas are included in the dataset. The second scatter plot measures the number of bedrooms on the x-axis and the number of bathrooms on the y-axis. There is a general positive relationship, as expected, but there are certainly some anomalies. Among them are a unit with 8 bathrooms and 0 or 1 bedroom, and a unit with 10 bedrooms but only 1 bathroom.

1. Can you make a time series plot?

No, this is a cross-sectional dataset.

Insights from above analysis:

* Fairly even geographical split from east to west.
* 4 clusters of units (metropolitan areas) on the east coast as compared to 2 on the west coast
* Review scores are not normally distributed, but prices are.

Correlations

* There is a positive correlation between bedrooms and bathrooms

Hypothesis to investigate further

* Pricing inefficiencies, when compared to how units are reviewed, mean that trusting bargains can be made on the consumer side. Conversely, it means that not all hosts are charging the right price based on reviews. This needs to be refined to compare only the groups of reviews and prices for a given number of rooms and bathrooms.
* I would like to investigate which city is on average the most expensive.

Narrative:

* Air Bnb hosts in metropolitan areas are concerned primarily with their unit being booked, not with charging the most accurate price possible. They would rather guarantee a consistent baseline of income than risk pricing incorrectly and having an empty unit. Economically speaking, they seek profit with a high quantity and a low cost as opposed to the opposite. In theory, this means that Air Bnb units as a whole a substitutable for each other. However, based on review score ratings, these hosts can afford to move the price up a bit without sacrificing quantity, and therefore increasing revenue.