# Capstone Statistical Inference Report

This report describes statistical tests performed on my capstone project dataset and their results. I performed more tests than those described here, but only the important ones are discussed. Full code is available on my GitHub page here: https://github.com/jkarpen/Springboard_Projects/blob/master/Capstone/Scripts/Capstone_Statistical_Inference.ipynb

Before starting testing I outlined the following questions I wanted to answer:

- Does Price follow a normal distribution? What if I convert to a log scale?
- How correlated are Beds, Bedrooms, Bathrooms, and Accomodates with Price? What about their correlation with Utilization?
- Are Price and Utilization correlated?
- For hosts that respond 100% of the time, are any of the response time groupings more related to price than the others? What about Utilization?
- For each Amenity, how does it impact Price/Utilization (using t-tests against the rest of the listings that don't have that amenity).
- For each Property Type, how does it impact Price/Reviews per Month?
- Using the Yelp data, how are number of businesses, number of reviews, and avg. star rating correlated with Price/Reviews per Month?
- Using the Yelp data, is there a significant Price difference between listings with no businesses w/in .1 mile and those that do? What about between those that have no businesses w/in .5 mile and those that do? Same for Reviews per Month.

Before starting I had to calculate Utilization Rate, because it is not part of the original dataset. I used the formula on pages 49-51 of this document from San Francisco’s Budget and Legislative Office in a report discussing the impact of short term rentals on the city: https://sfbos.org/sites/default/files/FileCenter/Documents/52601-BLA.ShortTermRentals.051315.pdf

# Test Results

## Beds/Bedrooms/Bathrooms/Accommodates Correlations

As suspected based on the charts created for the Data Story Report, Price shows a fairly strong correlation with all four of these variables. The correlation coefficient is .58 for Accommodates, .48 for Beds, .44 for Bedrooms, and .3 for Bathrooms. All have a p-value close to 0 indicating they are statistically significant.

Utilization, on the other hand, does not show a strong correlation with any of these variables. The largest coefficient is .1 for Accommodates, with a p-value very close to 0. The other coefficients are around 0.

## Which Amenities have the greatest impact on Price?

Previously, as part of the Data Wrangling phase, I unpacked the Amenities variable into multiple flag variables, one for each specific Amenity. Now I looped through the flag variables and for each Amenity, split the data frame into listings that have it and those that do not. I then performed a t-test comparing the mean Price of each group.

The most impactful Amenities that lead to an increase in Price are having a TV (good for $60 higher price), a pool ($36), a gym ($20), and Wifi ($13).

There were many more Amenities that contributed to a decrease in price. The ones with the biggest impact are having Free Street Parking (good for a $62 decrease - maybe because it means a lack of guaranteed parking), a Dryer ($54), a Washer ($50), Pets Living on Property ($47), and Smoking Allowed ($47). Interestingly, listings that say they have Heating also saw a $44 decrease in mean price, which is surprising for an area that gets as cold as Toronto. Only 1,741 out of 16,000 listings advertise Heating, so perhaps most listings take it for granted, and the ones that advertise it don't have much else going for them.

## What is the impact of Property Type on Price?

I performed a similar process on Property Types, using t-tests to compare each property type against all the others.

The property type with the largest mean Price increase are Serviced Apartment ($56 increase), though only 163 of those exist in the dataset. Also Condominium ($43) and Loft ($40). On the negative side, Guest Suite ($36 decrease), Bungalow ($40), and Guesthouse ($42).

## What is the impact of having a restaurant or bar within .1 and .5 miles of a listing?

Next I divided the dataset between those that have a restaurant or bar within .1 mile and those that do not, and did a t-test to compare the mean price of the two groups. Having a business within .1 mile is good for a $35 mean Price increase, with a p-value extremely close to 0 indicating this is statistically significant.

Having a business within .5 miles leads to a $31 price increase, and this relationship is also statistically significant.

## How are the number of restaurants/bars, total reviews, and avg. rating correlated with Price and Utilization?

Looking at the Yelp metrics within .1 mile of a listing, there was only a small correlation against either Price or Utilization. None of the correlation coefficients were higher than .089, though all p-values were small enough to show statistical significance.

A different picture emerges when looking at the data for businesses within .5 mile of each listing. Both number of businesses and total number of reviews for those businesses show a .23 correlation coefficient against Price. Average rating shows a very small negative correlation at -0.06. Utilization does not show a strong correlation with any of these variables.