Major Findings

The dataset appeared very clean already, but we began by checking for duplicates and empty/NaN entries, just in case. After doing this we began by making some initial plots of the data, these initial plots all appeared very odd though, with one group towering over the rest in most of the bar charts. So, to check for outliers, we generated a boxplot, which is pictured below. As we had suspected, there were quite a few outliers, with one especially being far, far greater than every other value. So, we then made a new clean dataframe, containing only rental prices below the upper bound.

Note: We assume Rental Price is measured in Indian Rupees in this dataset, but the original author on Kaggle unfortunately doesn't specify the units for it anywhere, so we can't be 100% sure.

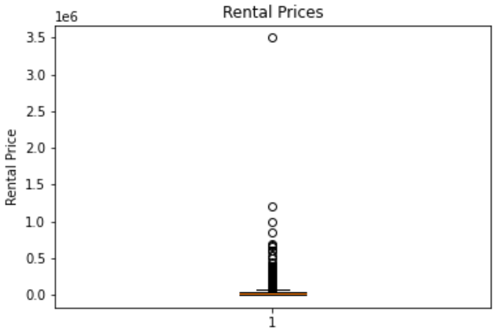
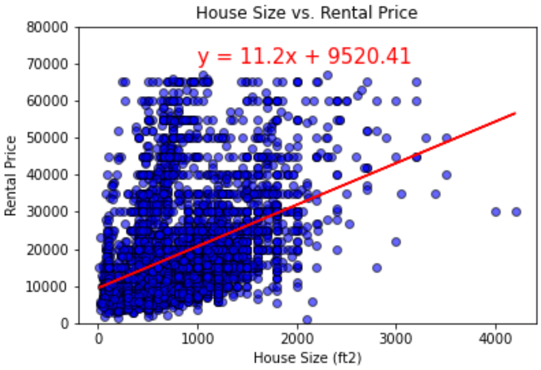


Figure 1. Rental Price boxplot.
Values above 67500 may be outliers.

Does the house size affect the rental price?



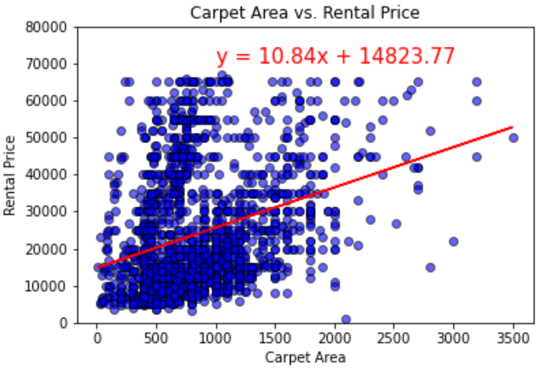


Figure 2. House Size vs. Rental Price.
R=0.394, R^2=0.155.

Figure 3. Carpet Area vs. Rental Price.
R=0.325, R^2=0.106.

Hypothesis: House size affects rental price.

Yes, it appears that house size does indeed have some effect on rental price, though the correlation isn't as strong as we had predicted. Figure 2 has an r-value of 0.394 and an r-squared value of 0.155. This r-value indicates a moderate positive correlation, though not a strong one. The r-squared value indicates that house size is responsible for approximately 15.5% of the variation in rental price, this value is lower than expected, but it is not totally unsurprising, as we would expect house size to be only one of several factors that influences rental prices.

Figure 3 is a similar graph, however house sizes measured via super area and built area (Area Type definitions explained in the Glossary) have been removed. We suspected that super area in particular, may have been skewing the data, as super area also includes common areas (lobbies, elevators, etc.) in its size measurements, meaning that these properties may appear very large based on house size measurements, but potentially not actually be very expensive. However, despite removing both super area and built area, the scatter plot and the correlation appears mostly unchanged, with an r-value of 0.325, and an r-squared value of 0.106.

The hypothesis does seem to be supported by this data, but not to the extent we had expected.

Does the area type affect the rental price?

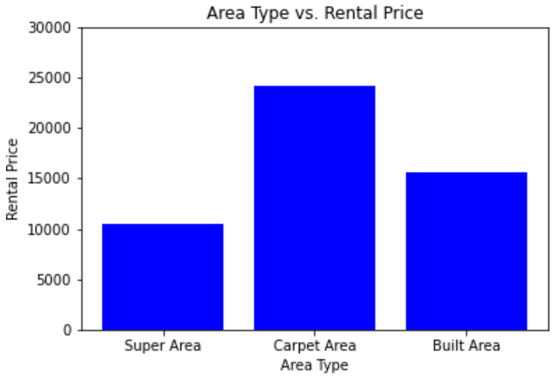


Figure 4. Area Type vs. Rental Price.
P<0.0001

Hypothesis: Super Area rentals have a lower rental price than others.

Yes, it appears that area type also affects rental price. A one-way ANOVA returned a P-value of 1.376e-93, a very low value, well below the 0.05 threshold of statistical significance. However, built area has an extremely small sample size (n=2) relative to the other groups, so we then performed a T-test just comparing super area to carpet area. This T-test also returned a P-value below 0.0001, confirming statistical significance and allowing us to confidently reject the null hypothesis here. It appears that properties measured by carpet area tend to be the most expensive, with super area being the least expensive. This makes some logical sense, as super area is inclusive of common areas, so it is likely that these tend to be smaller rooms/properties, potentially located in large apartment complexes or similar buildings. So, in regard to the hypothesis, the data appears to support it.

Does city/location affect the price?

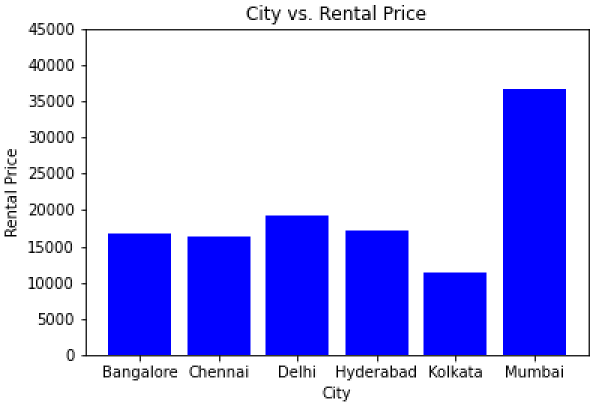


Figure 5. City vs. Rental Price.
P<0.0001.

Hypothesis: City affects rental price.

It also appears that city has an impact on rental prices. Similarly to above, we performed a one-way ANOVA which returned a P-value of 2.878e-309, an extremely low value. Based on this we can reject the null hypothesis here, in favour of the alternative hypothesis. It appears here that Mumbai is the most expensive city for rental properties by quite a large margin, but further testing is required to confirm this. Use of post-hoc analyses such as Tukey’s test or Bonferroni’s test would be useful in this regard.

Does the furnishing status affect the price?

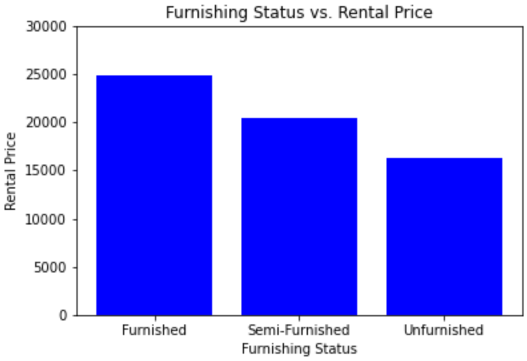


Figure 6. Furnishing Status vs. Rental Price.
P<0.0001.

Hypothesis: More furnished properties are more expensive.

Again, the P-value obtained from a one-way ANOVA (P=2.049e-39) indicates statistical significance and supports the hypothesis. Allowing us to reject the null hypothesis. Here, it seems that, on average, the more furnished the rental property is, the more expensive it will be, which makes logical sense, providing further support to the hypothesis. However, as previously stated, further testing is needed to confirm or refute this.

Does the point of contact affect the price?

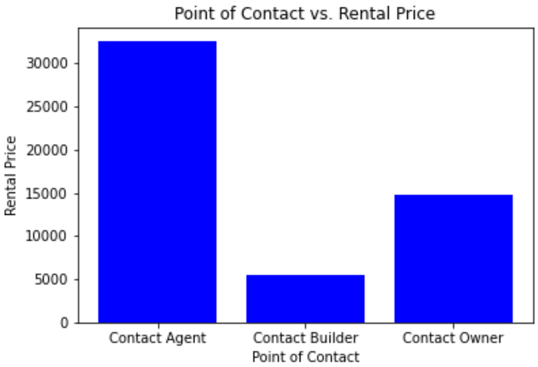


Figure 7. Point of Contact vs. Rental Price.
P<0.0001.

Hypothesis: Contact agent rentals will be more expensive.

Point of contact seems to affect rental prices, as P=0***.** As P<0.0001, we can confidently reject the null hypothesis here, as the results support the alternative hypothesis. It appears that rentals listed with their point of contact as "contact agent" tend to be by far the most expensive, while properties where the point of contact is the owner are far cheaper, though further testing is needed. Conclusions shouldn't be drawn from the "contact builder" column, as it has a sample size of n=1.

***** This obviously can't be an accurate P-value, but after much googling and discussing this odd P-value with the tutors, the most likely conclusion we found is that Python has a cut off for P-values, where P-values below a certain threshold will just appear as 0. To check that this wasn't being caused by the low sample size of the Contact Builder column, we did also do a T-test between Contact Owner and Contact Agent, which also returned P=0.

Does the number of bathrooms affect the price?

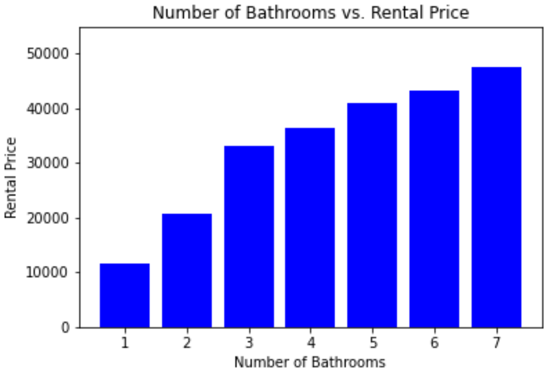


Figure 8. Number of Bathrooms vs. Rental Price.
P<0.0001.

Hypothesis: Rentals with a greater number of bathrooms are more expensive.

As P=4.209e-277 (obtained from one-way ANOVA), there seems to be a statistically significant difference here too, indicating that number of bathrooms does impact rental price. The trend observed here is also in line with expectations, as rental price seems to increase proportional to the number of bathrooms, supporting the hypothesis and, along with the P-value, allowing us to reject the null hypothesis.

Does the population affect prices?

Hypothesis: Population affects rental price.

Population numbers for each city were obtained from the population API found on api-ninjas.com. A graph wasn’t made here as it would look identical to the City vs Rental Prices graph (Figure 5). A one-way ANOVA here allows us to reject the null hypothesis again, with a P-value of 2.878e-309.

Conclusion

In conclusion, it appears that all the variables examined impact on rental prices. House size was not as important as we had expected prior to this analysis. Rental properties measured by carpet area generally seem to be more expensive than others, while ones measured by super area are the cheapest, likely due to super area properties, since super area includes common areas, generally being smaller properties located in apartment complexes and such. It appears that rentals in Mumbai tend to be more expensive on average than those in other cities. As expected, the more furnished a property is, the greater the rental price is likely to be, based on these results. Rentals with the point of contact listed as “contact agent” seem to usually be far more expensive than those listed as “contact owner”, likely due to the agent taking a cut, and the number of bathrooms appears to be positively correlated with rental price. Though, in all of these cases, further testing is required to confirm these findings. Post-hoc analyses such as Tukey’s or Bonferroni’s test would likely prove useful in providing further insight into the categorical variables here (the data presented as bar charts).