Zameen Property Data

Pakistan's realty state has a large contribution in its economic growth. According to the World Bank estimate, the size of a country's real estate assets constitutes between 60 and 70% of the country's total wealth; if these estimates are applied to Pakistan, the estimated size of the real estate sector would be $300 to $400 billion [1]. With the rate of urbanization that Pakistan has been experiencing, there is a growing need for urban planning [2].

In this project, we attempt to understand the needs of the real state in Pakistan, across different cities of Pakistan for properties which are on sale and available on rent.

About Dataset:

The Dataset was of property advertisements posted from August 2018 to July 2019 on a real-estate property website called zameen.com. The data-set can be accessed on Kaggle here.

Preprocessing:

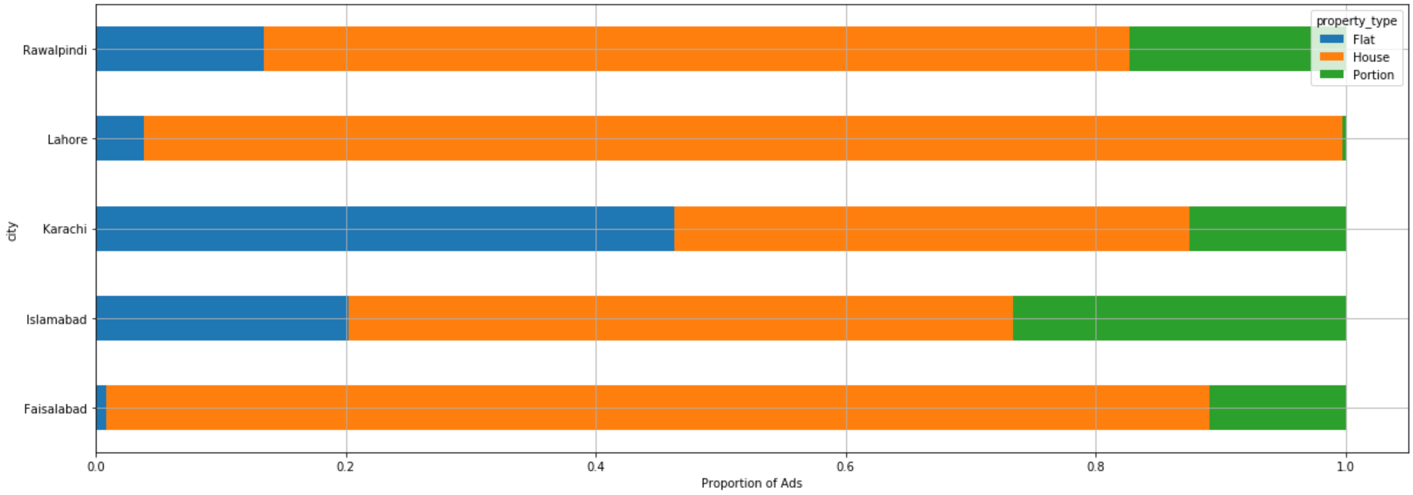
The dataset had advertisements posted by both real-estate agents and by individuals. However, the advertisements posted by individuals were as NaN so their value was replaced by the string 'Individual'. Furthermore, the dataset had both marla and kanal as metrics for property size. To be actual of size for measurements, one should stick to one measurement and a very common measurement metric in Pakistan is square yards [3] so we shifted property size to square yards. Our data set had the metric of location, however some locations were not unique. For example, DHA society is in Karachi and Lahore. Therefore, we introduced another attribute by combining city and location called unique location. To understand price and area together, we introduced another attribute for price per area. For outlier detection we calculated z-score on price per area separately on property for rent and for sale.

Questions:

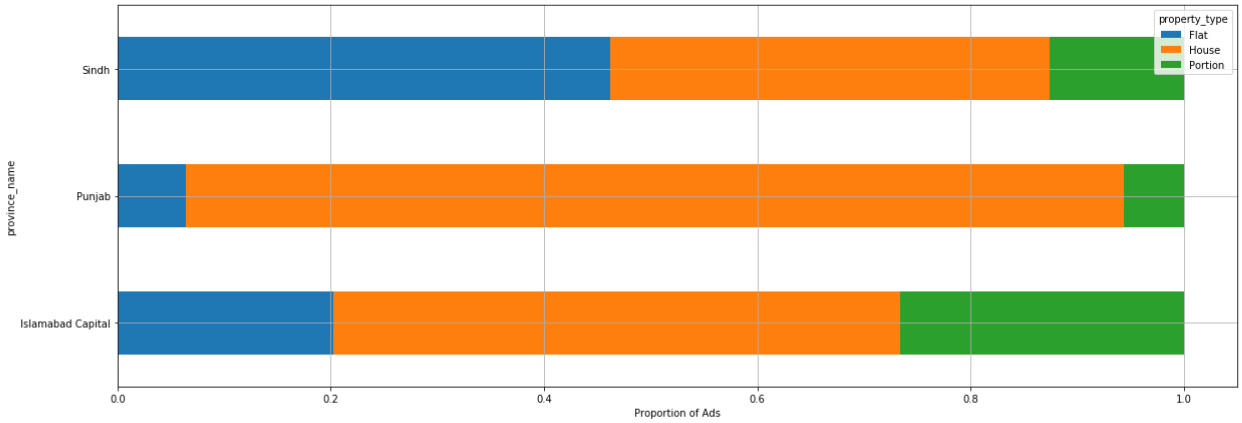
Question 1:

Explore where there exists any difference in property type across cities or provinces?

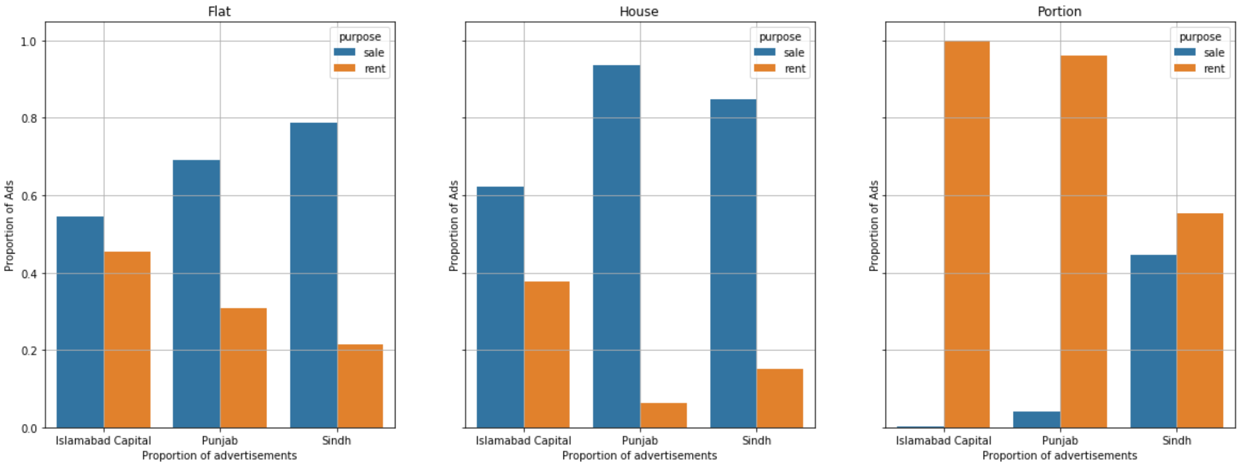
For this question, we explored the property across each city. We normalized the property type across each city such that the proportion of a certain property type only the proportion only within that city. We did so because we had a low number of advertisement posts for certain cities.



Here we see that there is a significant difference in proportions of advertisements. We see that Karachi has a significantly higher portion of flat, while Islamabad has a higher portion of property type. Generally speaking, Houses are the most common property type. We collapsed the data to provincial level and we see a similar pattern.



The data on provincial level follows a similar pattern. We went ahead to analyze if any property type is posted more frequently for a certain purpose.

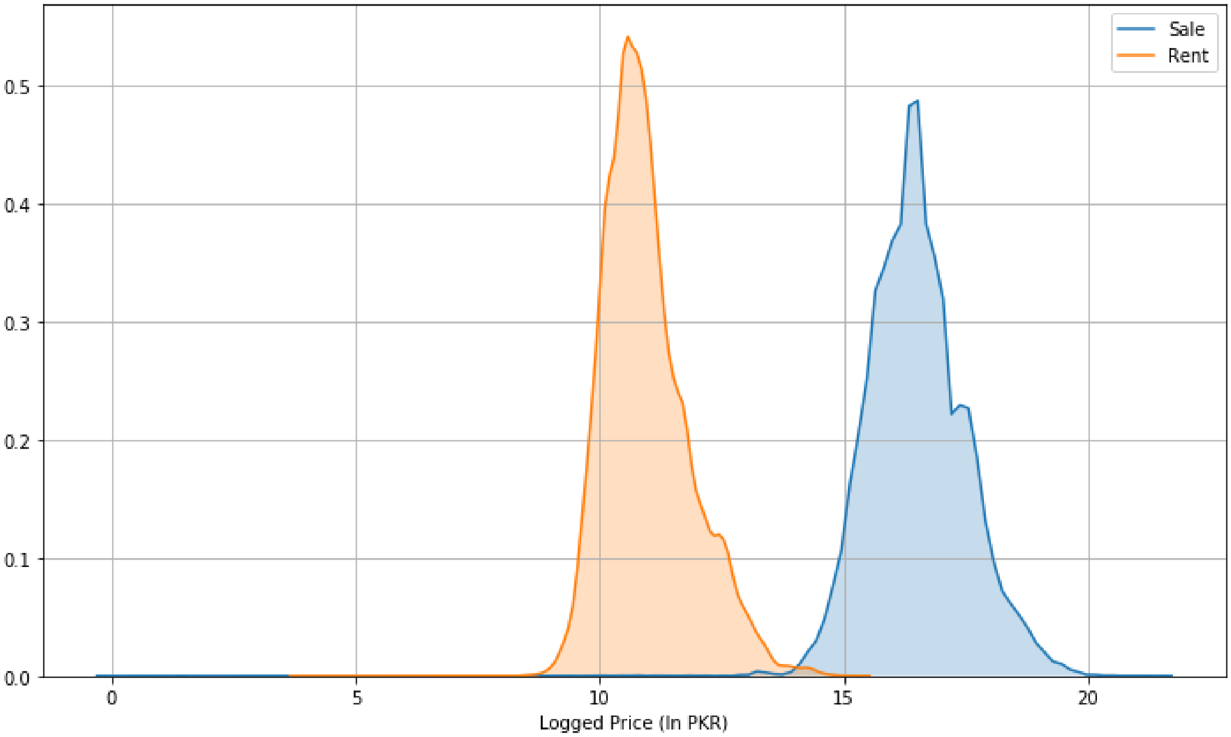


We see that usually most properties have greater proportion for the sale. Interestingly though, portions are offered more for rent than sale.

Question 2:

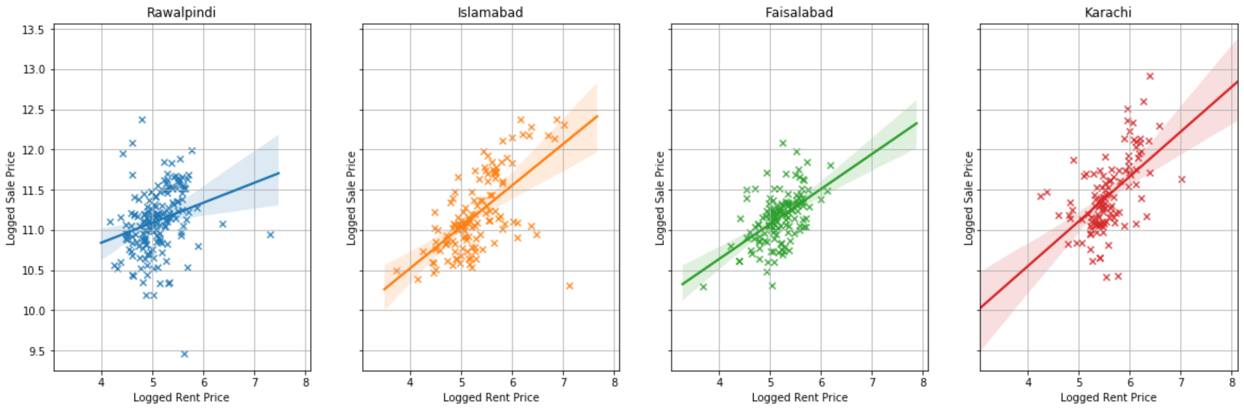
What is the relation between sale and rent in terms of recovery period? By recovery period we mean that the time period in which rent becomes equal to sale.

For this question we first analyzed the distribution of rent and sale prices.

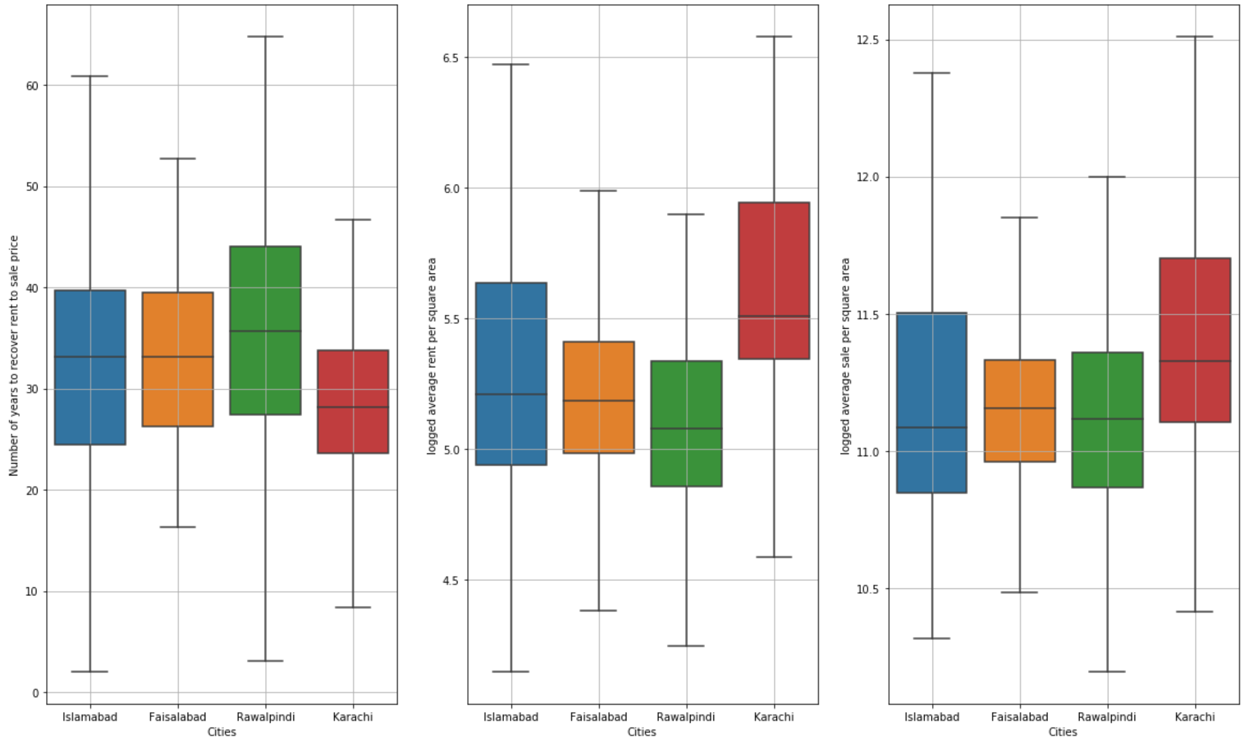


We see that both the distributions appear to be following a narrow normal distribution.

To compute the recovery period, we grouped the properties across each location and computed the mean rent per sq. area and sale per sq. area to try and compute the relation between rent and sale in that location. Looking at mean sale and rent per area for each city provides us with this figure.



We see that in some places, like Rawalpindi, the rent and price does not appear to be following a linear relation, as compared to Islamabad where rent increases with sale price. Karachi appears to be the extreme case having higher mean sale and rent price and following the steepest linear relation.



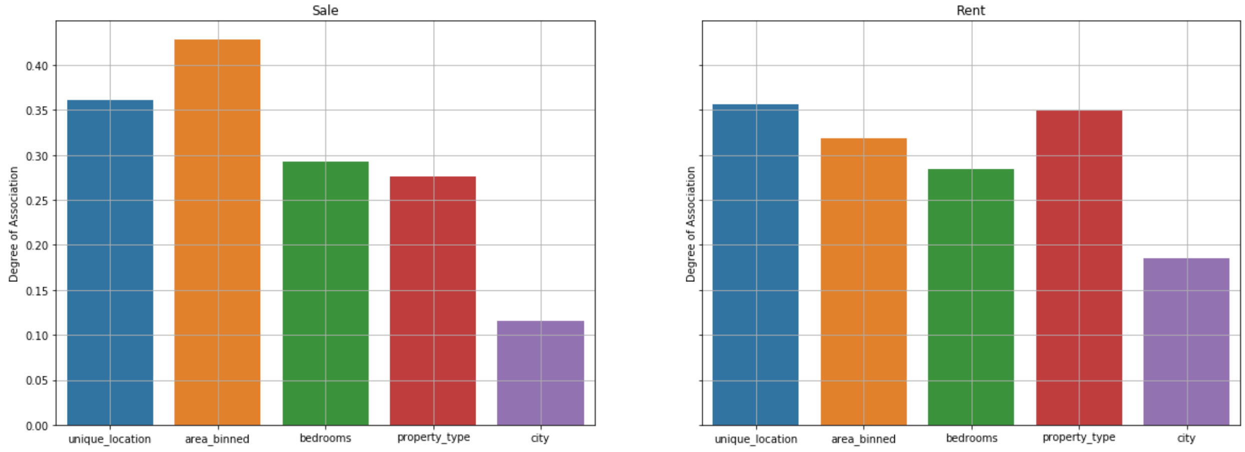
This figure extends the analysis further. We see that Karachi has a much higher rent per sq. area but not as high sale price which translates into the smaller recovery period, and hence in Karachi it takes approximately 25 years to recover while in other places it takes 35 years. It is interesting to see the similarity between Islamabad, Faisalabad and Rawalpindi. Karachi hence appears to be the most expensive city to live on rent.

Question 3:

On which factor does price depend on: location, area, number of bedrooms, property type?

For this question, we binned the numerical features into equal sized bins. We decided to go for equal sized bins as compared to equal width bins as we have a highly right skewed data and equal width bins result in extremely unbalanced bins. We did the binning in order to use measures for categorical correlation.

We used the prices computed for the last question to compute the correlation for this part. Upon using the simple chi-square test we were getting p-values = 0 for every case. Research suggested that it is because we have a large n and p-values would approach zero even on the slightest divergence from null hypothesis. Which is why we use the Cramer's V test, which tackles this problem.

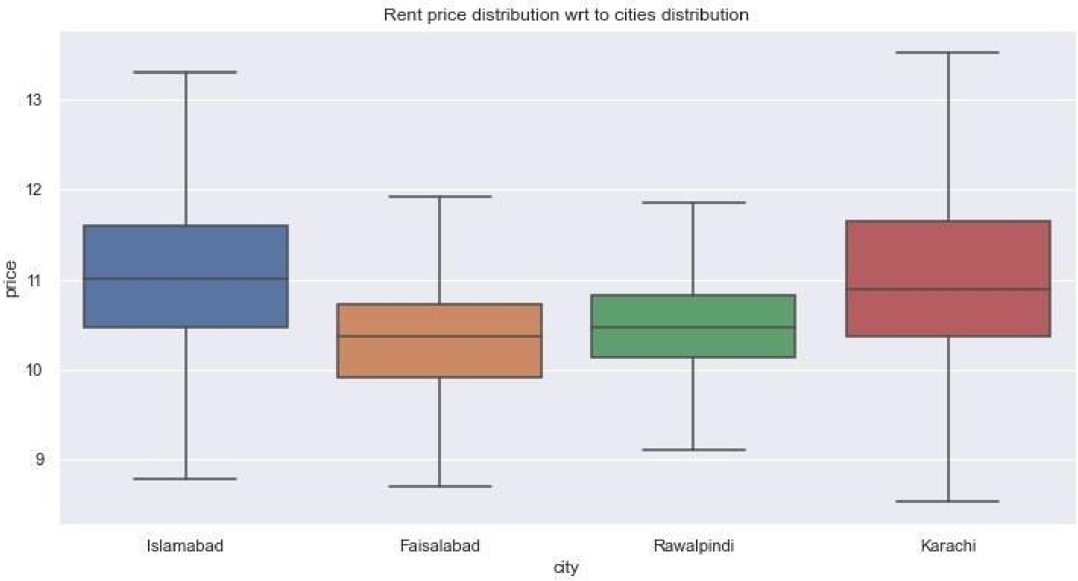


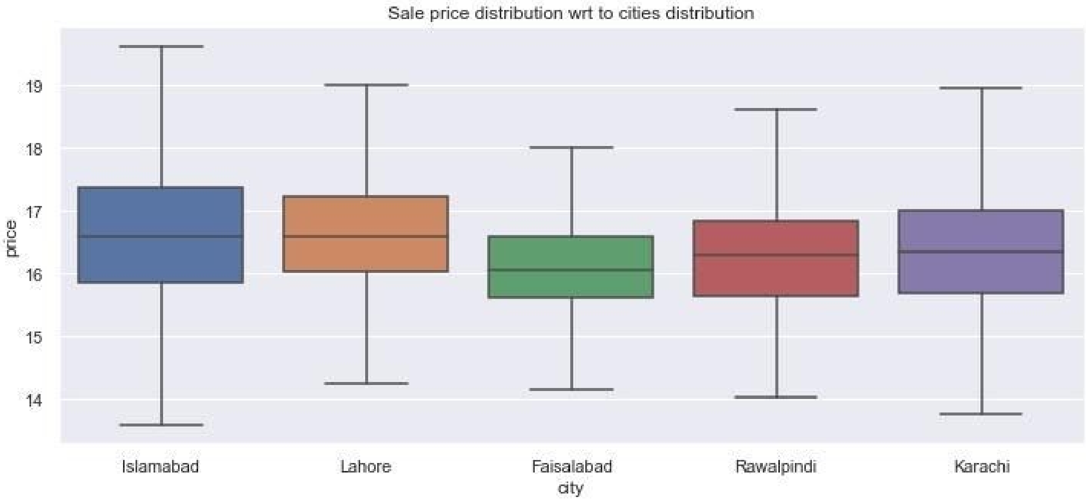
We see that for sale price, area and location appear to share the highest relation while in rent, we see that location and property type appear to share the highest relation which.

Question 4:

Difference between price across cities?

This relation has been explored earlier as well, but as a general measure, the plots below report distributions of sale and rent across cities.

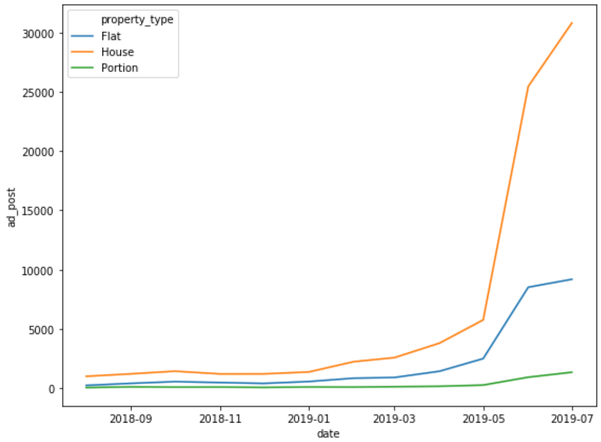


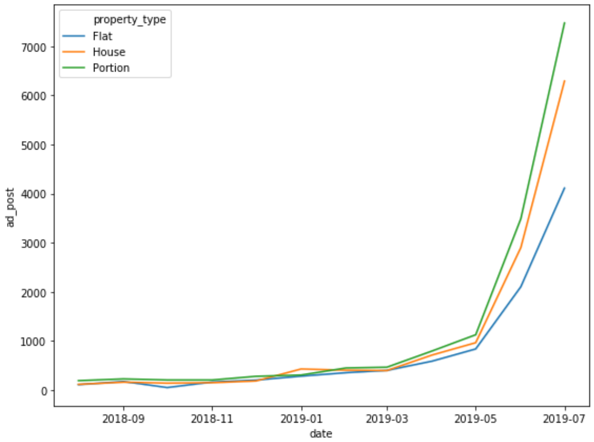


Question 5:

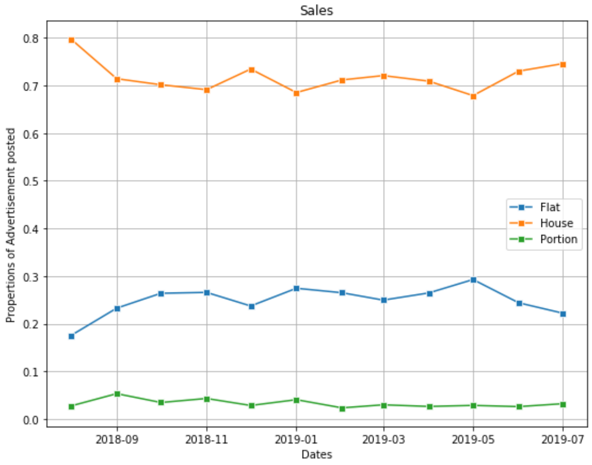
Trend of property-type with respect to time?

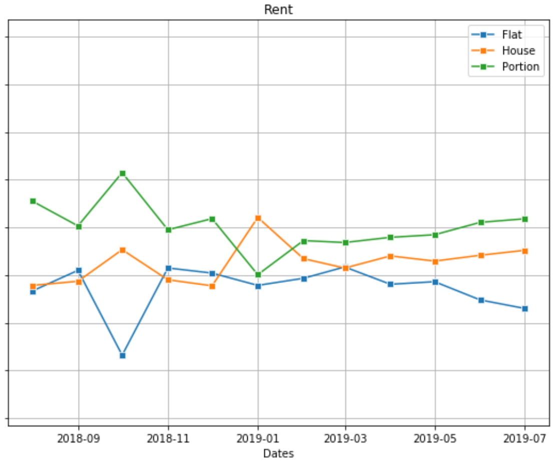
For this purpose we grouped all the records with respect to year and month of the date_added, we did so to smooth out the effect. Just to view, we plotted the number of advertisements posted.





We see that we have very small data prior to May 2019, and thus most of the time analysis would not be very accurate. Still, in order to analyze further we normalized the property type with results to total advertisements to see if there is any pattern.





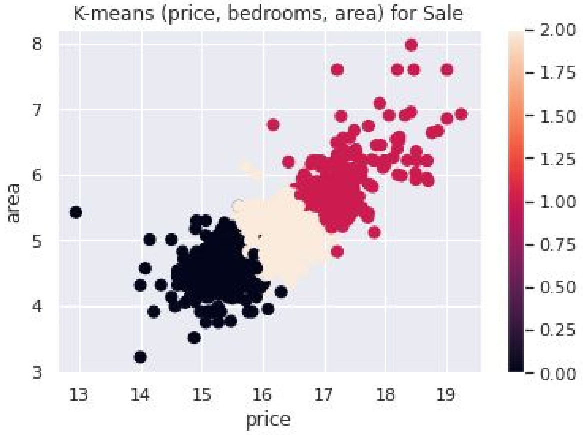
This is an interesting figure, we that for sale houses constantly share the greatest proportion. Approximately 70-80% of the ads are for houses and 20-30% of the ads are for flats across Pakistan. For rent, though, we see a mumbled and much more congested proportion, though we do see ‘portions’ to be dominating this as they share 40% percent of the proportions of advertisement posted while houses on the other hand share approximately 30%.

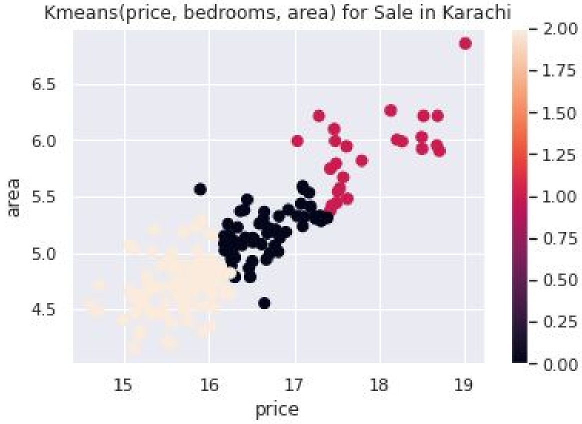
Question 5:

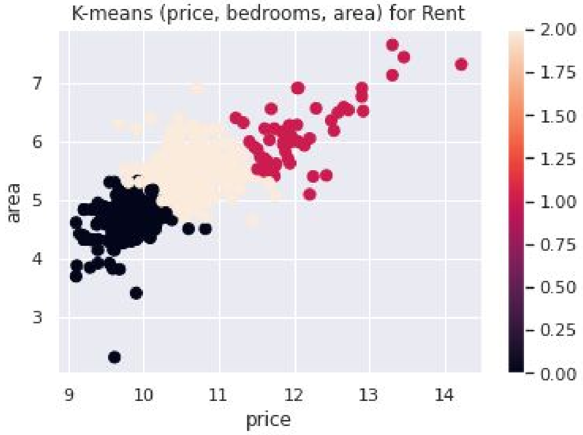
Clustering analysis of attributes overall and across cities.

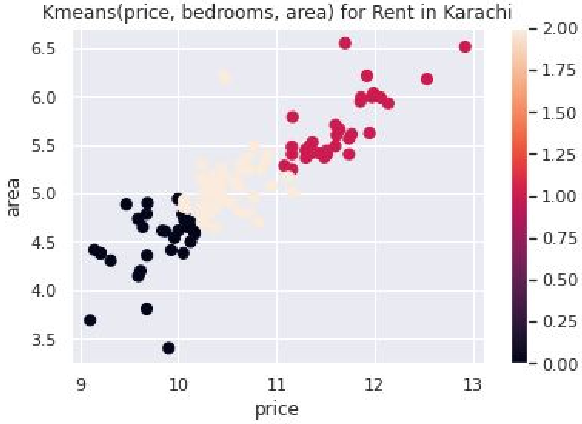
For this purpose, we tried three clustering approaches: k-means clustering, hierarchical clustering using dendrogram and Density-based spatial clustering of applications with noise (DBSCAN).

We clustered over the important numerical variables: prices, area and bedrooms, on sales and rent separately.

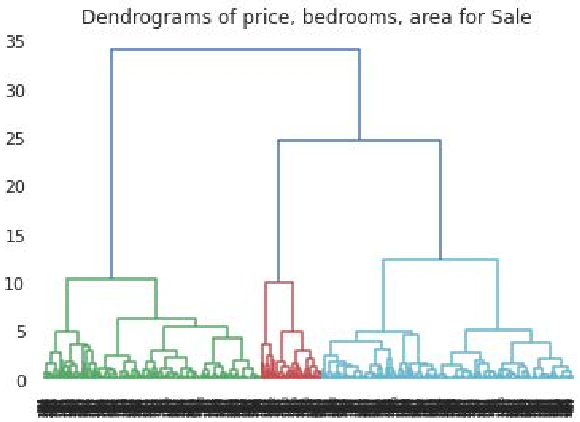

K-means (price, bedrooms, area) for Sale

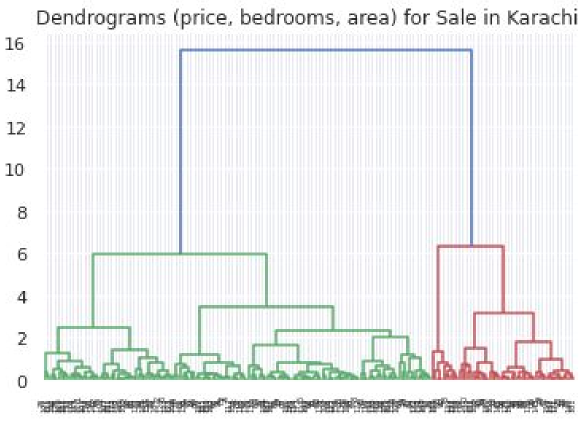

Kmeans(price, bedrooms, area) for Sale in Karachi

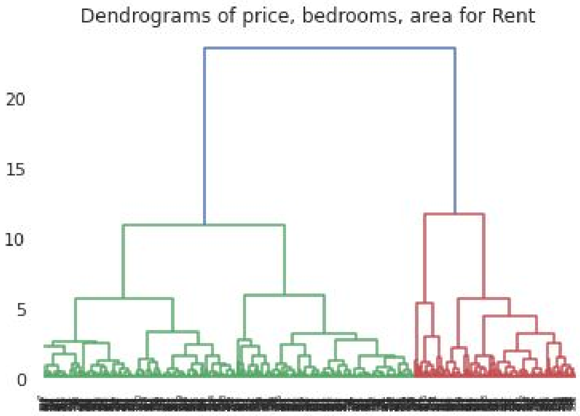

K-means (price, bedrooms, area) for Rent

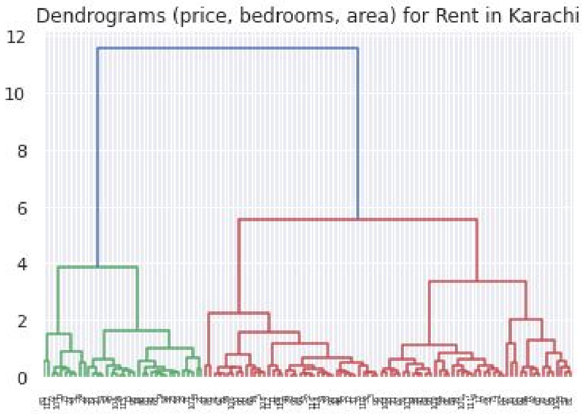

Kmeans(price, bedrooms, area) for Rent in Karachi

For k-means analysis we made 3 clusters. We can see from the figures for both sale and rent, as anticipated that houses with large areas have higher prices, houses with mid-level areas have mid-level prices and houses with smaller areas have lower prices. We can see this in all our cities.

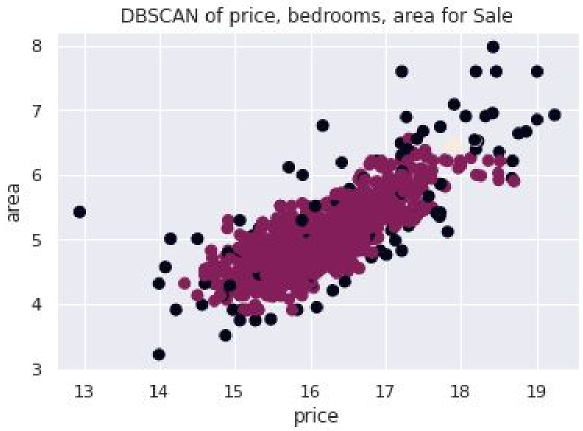

Dendrograms of price, bedrooms, area for Sale

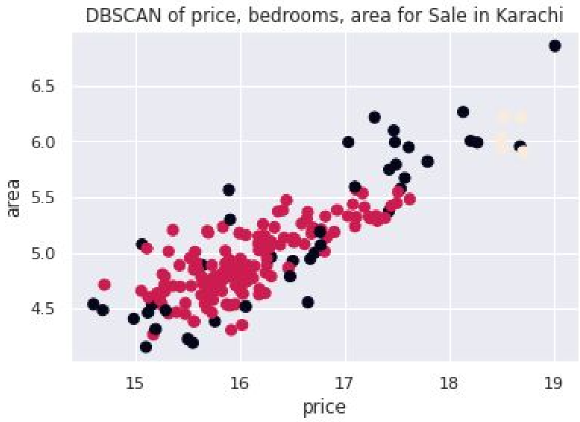

Dendrograms (price, bedrooms, area) for Sale in Karachi

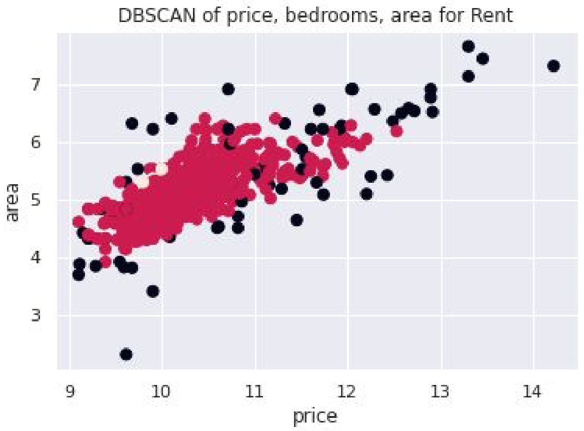

Dendrograms of price, bedrooms, area for Rent

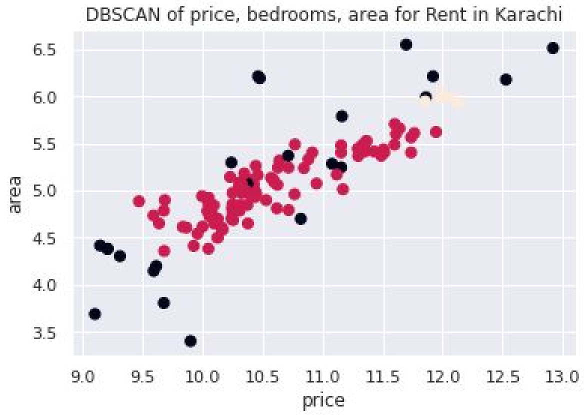

Dendrograms (price, bedrooms, area) for Rent in Karachi

From the dendrogram we can observe for both rent and sale that the difference between clusters is very huge as the height of the clades is varying. We also have a high number of observations available.


DBSCAN of price, bedrooms, area for Sale


DBSCAN of price, bedrooms, area for Sale in Karachi





The DBSCAN algorithm does not take the number of clusters so we get much more interesting using it. for sale of all cities, it makes 4 clusters with relation of price and area and for Karachi it only makes 3 clusters, which means that the price and area relation in Karachi has less categories than other cities. However for rent, it produces an equal number of clusters i.e. 3 for all cities and Karachi. We observe that all cities have more houses with mid-level areas and mid-level prices and very few small houses with low prices.

References:

[1] S. K. Atiq, “Pakistan’s 2020 real estate prospects and challenges,” *Profit by Pakistan Today*, Mar. 12, 2020. Retrieved from: https://profit.pakistantoday.com.pk/2020/03/12/pakistans-2020-real-estate-prospects-and-challenges/.

[2] D. Dowall and P. Ellis, “Urban Land and Housing Markets in the Punjab, Pakistan,” *Urban Studies*, vol. 46, pp. 2277–2300, Sep. 2009, doi: 10.1177/0042098009342599.

[3] “Kanal, Marla, Square Feet, Square Yards Conversion - Zameen.com Forum.” https://www.zameen.com/forum/discussions/other_and_misc/kanal__marla__square_feet__square_yards_conversion-12358.html (accessed Dec. 03, 2020).