EDA of Zameen.com

Libraries we used

- Numpy
- Pandas
- Seaborn
- Matplotlib
- Plotly

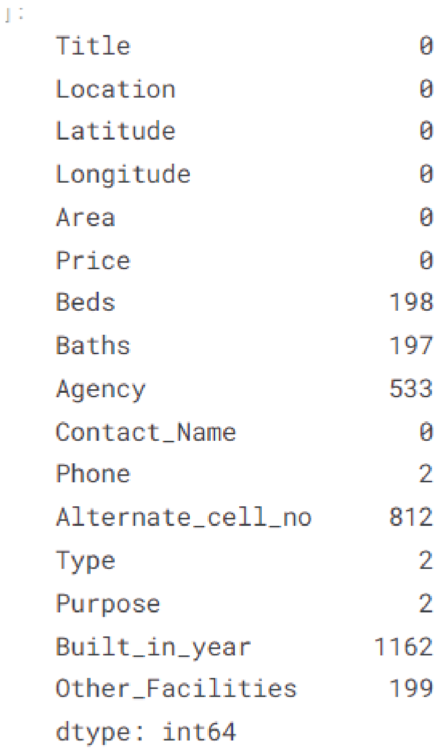
Steps involved in completing this project:

1. Import dataset

Import the dataset that I previously scraped from Zameen.com using python and its libraries like beautifulsoup, pandas and regex. Beautiful Soup is a Python package used for scraping HTML and XML documents. It creates a parse tree for scraped pages that can be used to extract data from HTML documents. Pandas is a Python library. It is used to manipulate and prepare data for analyzation. Regex or Regular Expression is a library that can be used to check if a string contains the specified search pattern.

2. Clean data

- First of all, I check null values present in columns of dataset using command:
  data.isnull().sum()



```
]:
Title                 0
Location              0
Latitude              0
Longitude             0
Area                  0
Price                 0
Beds                198
Baths               197
Agency              533
Contact_Name          0
Phone                 2
Alternate_cell_no   812
Type                  2
Purpose               2
Built_in_year      1162
Other_Facilities    199
dtype: int64
```

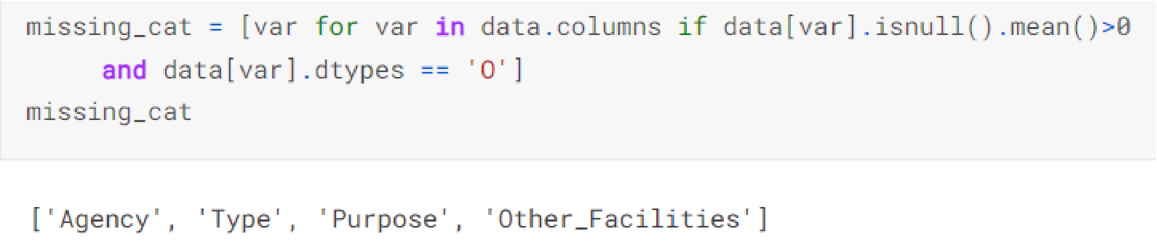
- I replace all null values from beds and baths column by median of all values of that column:
  median = data['Beds'].median()
  data['Beds'].fillna(median, inplace=True)

  median1 = data['Baths'].median()
  data['Baths'].fillna(median1, inplace=True)

- Then I convert all values of Beds and Baths column to integer, s that any further operation can be perform smoothly
  data[['Beds', 'Baths']] = data[['Beds', 'Baths']].astype('int')
- Null values in Phone and Alternate_cell_no columns where user have not provide any contact information is replace by 0.

  data['Phone'] = data['Phone'].fillna(0)
  data['Alternate_cell_no'] = data['Alternate_cell_no'].fillna(0)

- The remaining columns in dataset that have null values are categorical columns, these columns can be specified in any dataset:



```
missing_cat = [var for var in data.columns if data[var].isnull().mean()>0
     and data[var].dtypes == 'O']
missing_cat
```

```
['Agency', 'Type', 'Purpose', 'Other_Facilities']
```

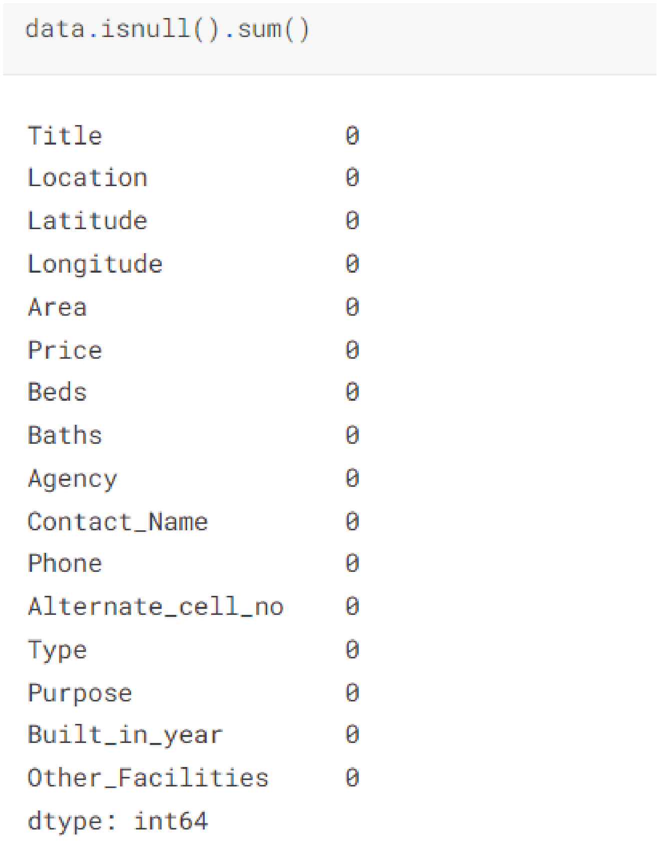
- In Type, Other_Facilities and Purpose columns null values are replaced by mode of total values in column
  data['Type'].fillna(data['Type'].value_counts().index[0], inplace=True)

  data['Purpose'].fillna(data['Purpose'].value_counts().index[0], inplace=True)

  data['Other_Facilities'].fillna(data['Purpose'].value_counts().index[0], inplace=True)

- In Built_in_year and Agency columns all null values are replace with word “Missing” and “Not Mentioned”as a lot of values from both of the columns were missing and any filling null values with any calculated value can produce inconsistent results
  data['Built_in_year'] = data['Built_in_year'].fillna('Missing')
  data['Agency']= data['Agency'].fillna('Not Mentioned')

- Now all null values from dataset has been replaced:



```
data.isnull().sum()
```

```
Title                0
Location             0
Latitude             0
Longitude            0
Area                 0
Price                0
Beds                 0
Baths                0
Agency               0
Contact_Name         0
Phone                0
Alternate_cell_no    0
Type                 0
Purpose              0
Built_in_year        0
Other_Facilities     0
dtype: int64
```

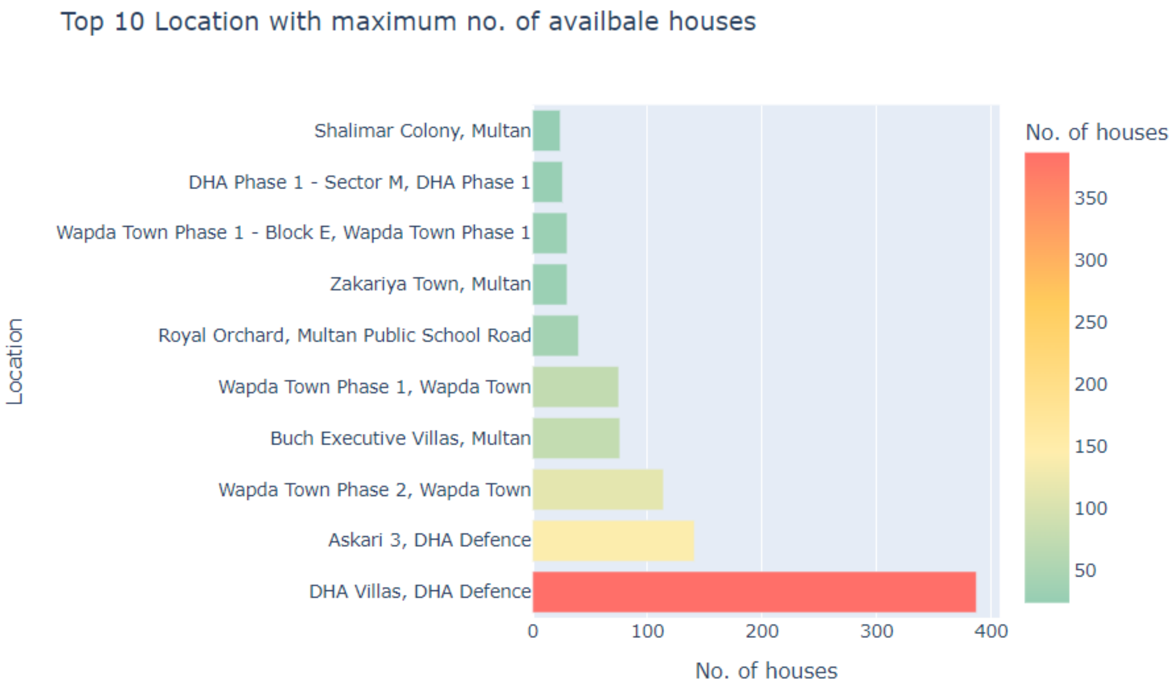
- Now add a column Price Bins according to the price ranges of house. Create bins for price range and then assign bins according to price range.

```
bins = [5000000, 28750000, 52500000, 76250000, 100000000]

price_bins = ['cheap', 'affordable', 'semi-expensive', 'expensive']


#make a Price_Bins column assign pins according to prices

data['Price_Bins'] = pd.cut(data['Price'], bins, labels= price_bins)
```

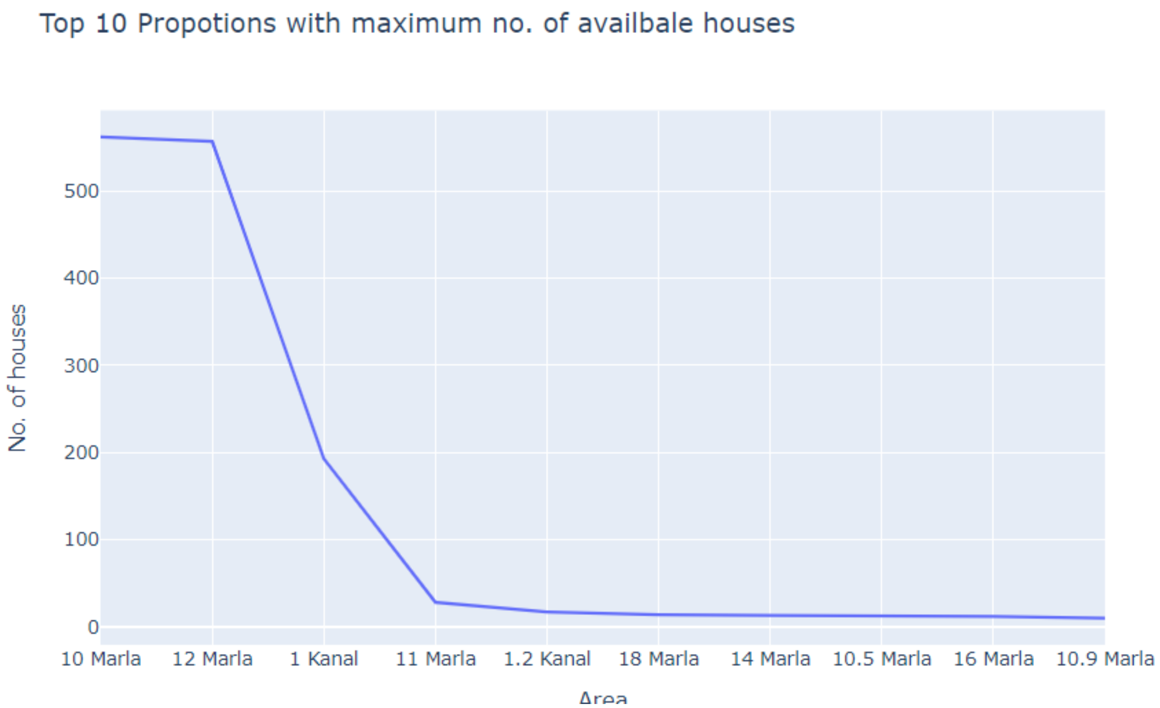
3. Perform EDA on data

1. The top 10 locations that have the most selling properties in Multan

I use group by function to group all properties on basis of their location and then count and sort all properties. Then I use Plotly express to plot graph and at x axis I plot No. of houses and at y axis I plot location of houses.

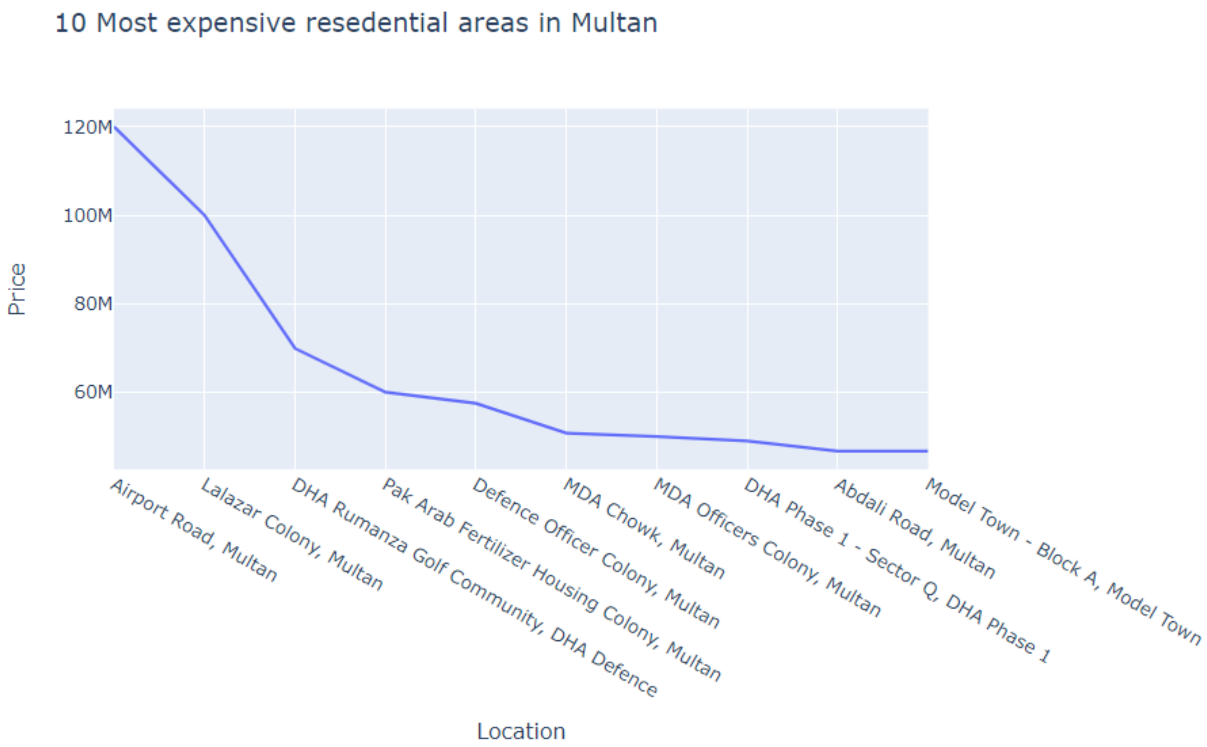


2. **In which area maximum no. of properties are available?**
   I use group by function to group all properties on basis of their area and then count and sort all properties. With Plotly express to plot graph and at x axis I plot the proportions of houses and at y axis I plot No. of houses.



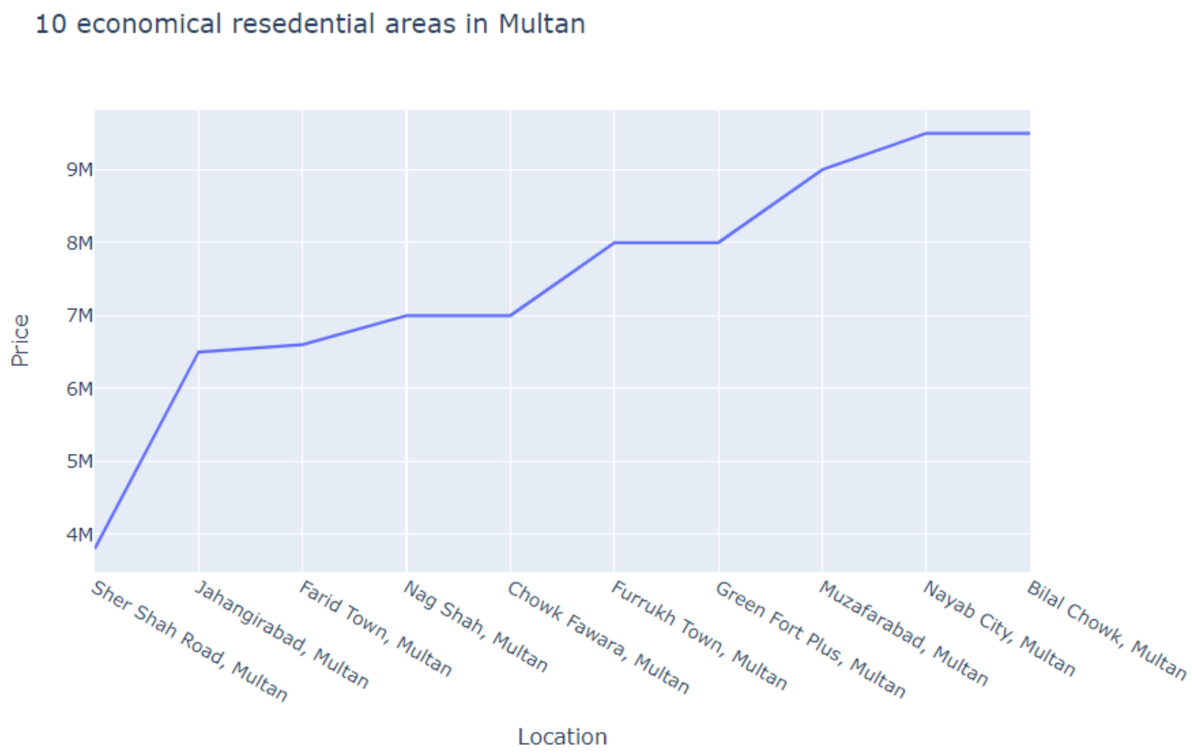
3. 10 most expensive resedential areas in multan?

I use group by function to group all properties on basis of their location and then take mean of prices and sort all values on basis of prices in descending order. With Plotly express to plot graph and at x axis I plot the location of houses and at y axis I plot average prices at that location.



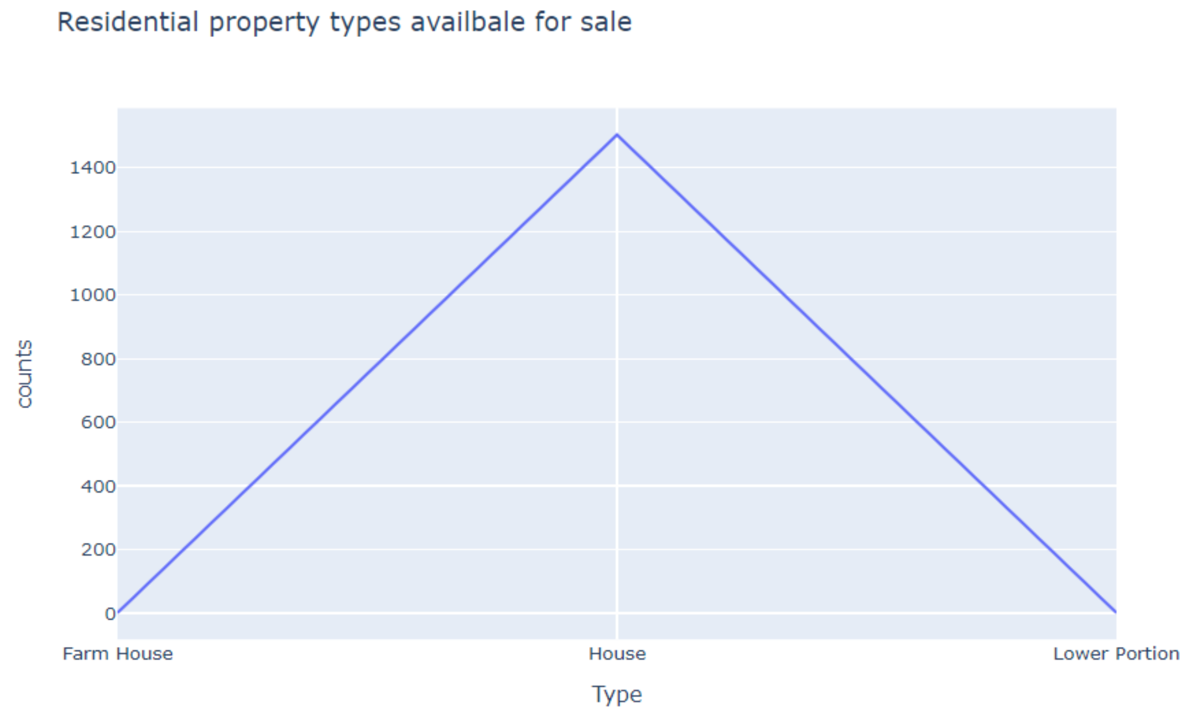
4. 10 most economical resedential areas in multan?

I use group by function to group all properties on basis of their location and then take mean of prices and sort all values on basis of prices in ascending order. With Plotly express to plot graph and at x axis I plot the location of houses and at y axis I plot average prices at that location.

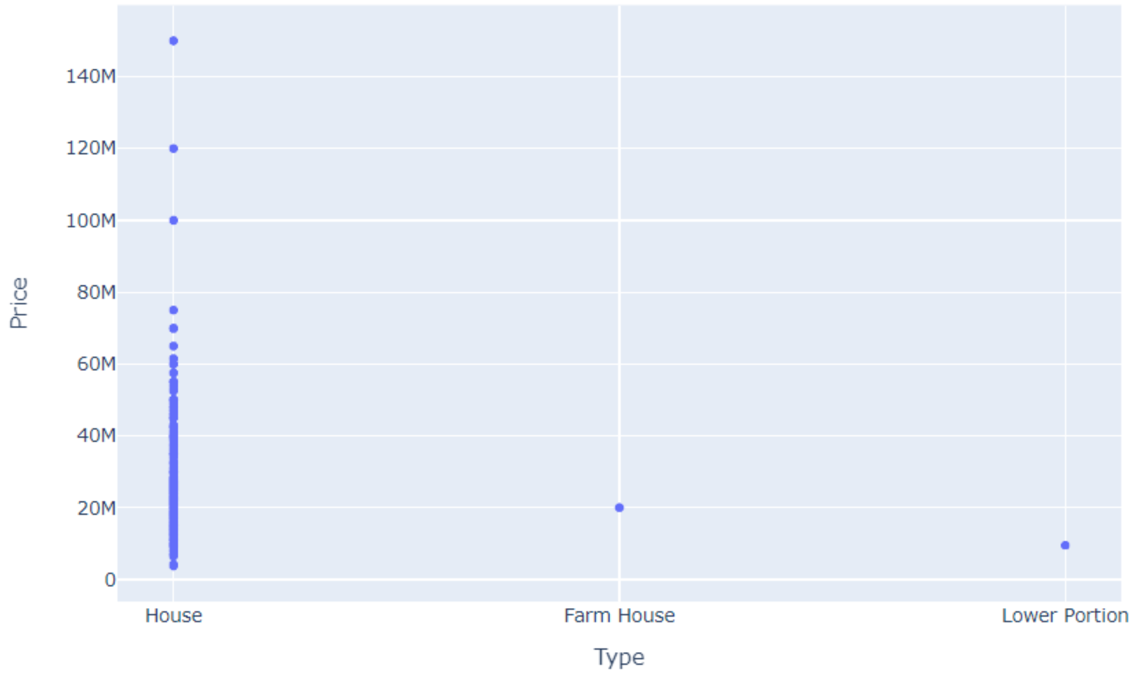


5. Most property type available for sale

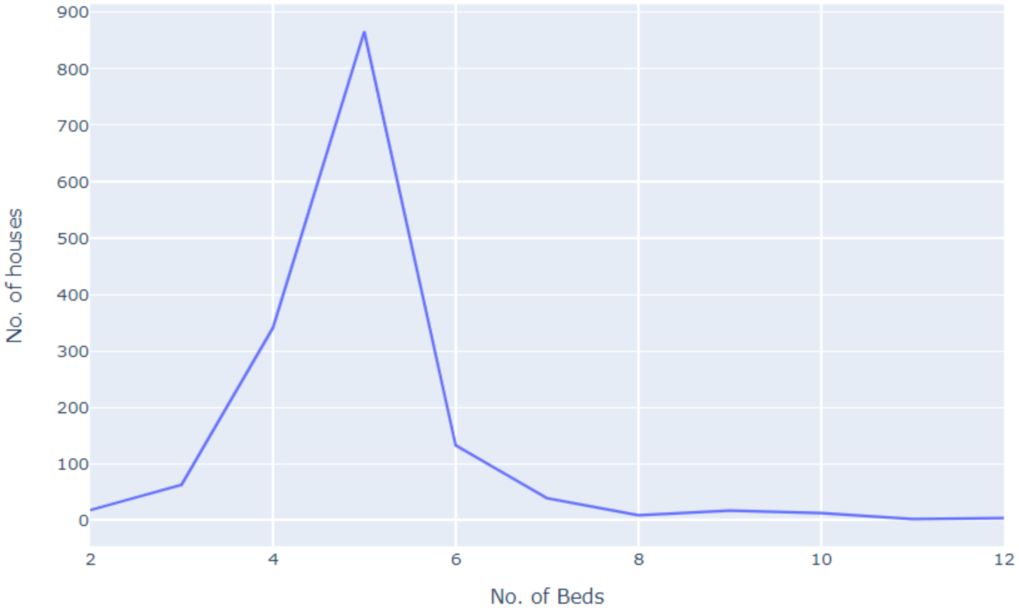
I use group by function to group all properties on basis of their type and then take size. With Plotly express to plot graph and at x axis I plot the type of houses and at y axis I plot No. of houses of that type.



6. **Price range for different types of houses?**
With Plotly express to plot graph and at x axis I plot the type of houses and at y axis I plot Price of houses

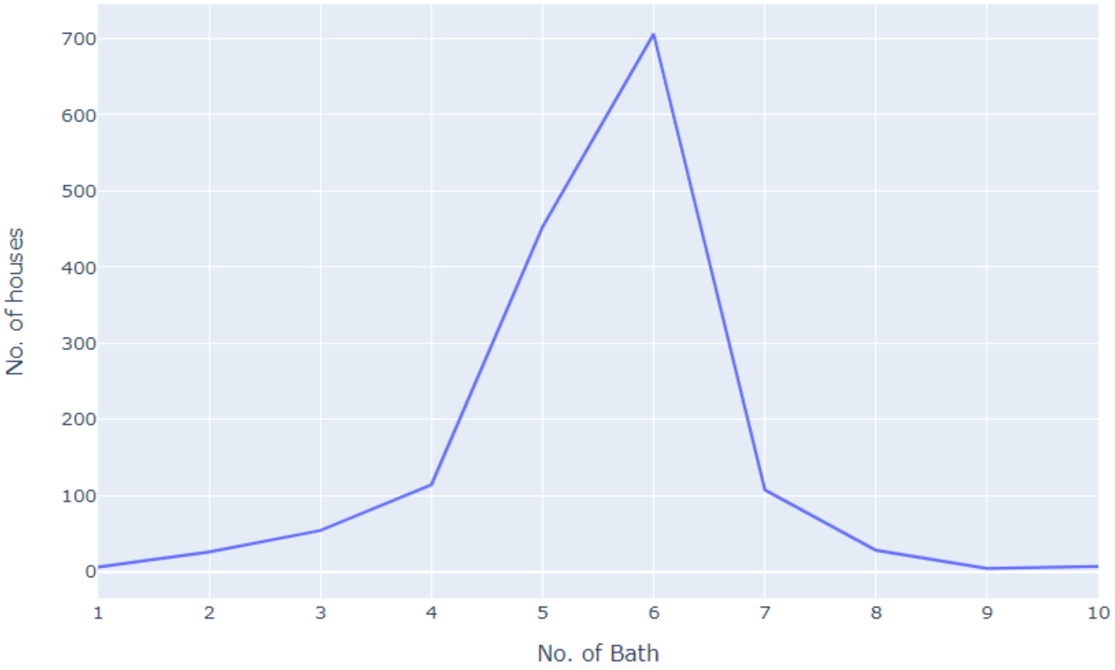


7. **Number of beds available in houses?**
   I use group by function to group all properties on basis of beds available in houses and then take size. With Plotly express to plot graph and at x axis I plot the beds in houses and at y axis I plot No. of houses of that type.

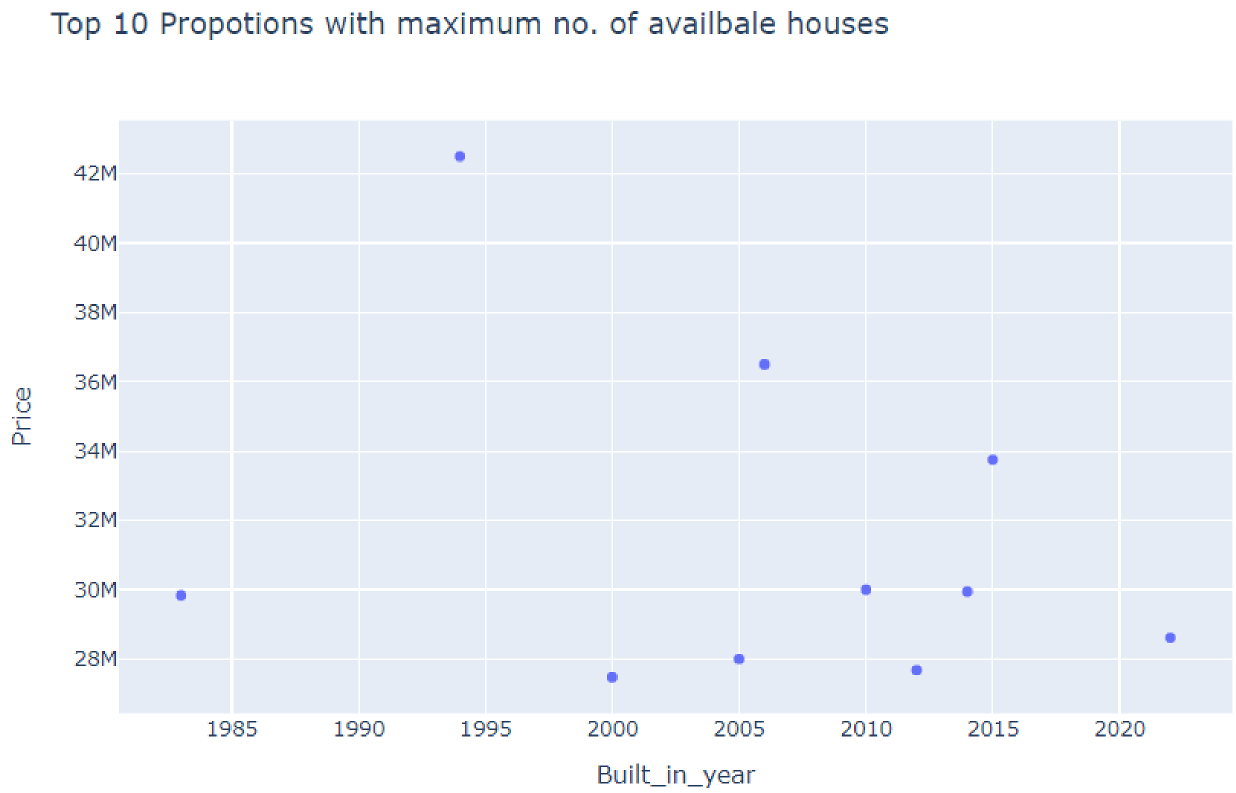


8. **Number of baths available in houses?**

   I use group by function to group all properties on basis of baths available in houses and then take size. With Plotly express to plot graph and at x axis I plot the baths in houses and at y axis I plot No. of houses.

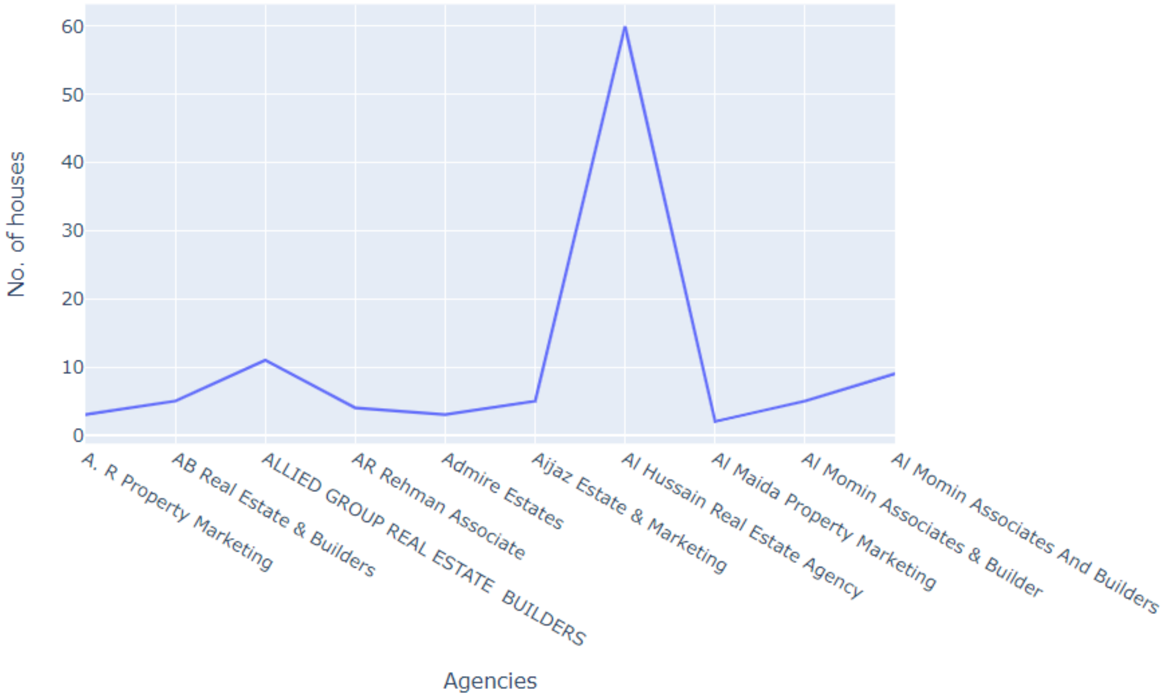


9. **Average price of houses in different years?**
   I use group by function to group all properties on basis of Built_in_year of houses and then take mean of their prices. With Plotly express to plot graph and at x axis I plot the Built_in_year of houses and at y axis I plot average prices of houses in that year.



10. Top 10 agencies that are selling most of the houses in Multan?

I use group by function to group all properties on basis of Agency that sell a house and then take size of it. With Plotly express to plot graph and at x axis I plot the real estate agencies and at y axis I plot No. of houses sell by that agency.



Results:

1. Most houses available are in 10 Marla, 12 Marla and 1 Kanal Proportions.
2. Results show that DHA Villas has a very large number of houses to sell
3. The Airport Road Area is the most expensive area to buy residential property.
4. Sher Shah Road, Multan is the most economical area to buy residential property.
5. Prices of House are in different ranges depending on area and location of house
6. Mostly Houses are available for sale only a few flats and farmhouses are also available
7. Mostly houses have 4-6 baths available
8. Mostly houses have 4-6 beds available
9. Al Hussain Real Estate Agency is most selling real estate agency in Multan