Varying King County House Price Analysis with respect to different Attributes

1. Overview of Data:

This dataset was acquired through Kaggle website, and the data was provided by King County, Seattle city in Washington. The Dataset deals with finance domain. The data includes details about the homes sold between 2014 and 2015 years. The dataset has 21 attributes. For description of attributes refer to Table Attributes.

Table Attributes

|  |  |
| --- | --- |
| Attribute | Description |
| id | House Number |
| date | House sold date |
| price | Price of each home sold |
| bedrooms | Number of bedrooms |
| bathrooms | Number of bathrooms, were  0.5 accounts for a room with a toilet but no bath,  0.75 accounts for a sink, toilet, bath, or shower.  1 account for a sink, toilet, bath and shower. |
| sqft\_living | Square footage of the apartment’s interior living space |
| sqft\_lot | Square footage of the land space |
| floors | Number of floors |
| waterfront | Variable for whether the apartment was overlooking the waterfront or not |
| view | On a scale of 4 the view of apartment is rated. 4 means Best |
| condition | On a scale of 5 defining the condition of the apartment, 5 means best |
| grade | On a scale of 1-13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11- 13 have a high-quality level of construction and design |
| sqft\_above | The square footage of the interior housing space that is above ground level |
| sqft\_basement | The square footage of the interior housing space that is below ground level |
| yr\_built | The year the house was initially built |
| yr\_renovated | The year of the house’s last renovated |
| zipcode | Area code of the house |
| lat | Latitude |
| long | Longitude |
| sqft\_living15 | On an average Carpet Area of the nearest 15 neighbors |
| sqft\_lot15 | On an average land lot area of the nearest 15 neighbors |

Reason to choose this dataset is, I found it interesting to analyze a house price based on multiple characteristics related to house irrespective of demographics. Dataset has vast scope for exploration as most of the attribute in dataset has moderate to high correlation with the price of the house. In-addition the columns describing the data are very realistic to understand. So, technically I thought it would be great for learning.

* Likely to answer

1. Can we assume the ideal house price based on the bedrooms and bathrooms?
2. Does the view of the house impact the price of the house?
3. Renovation of house has any relation to the grade of house.
4. Does Condition of the house varies depending on the year of construction, what could be the percentage of houses been sold with poor condition?
5. Overall, after how many years the houses are sold to get good returns?
6. Characteristic of the house sold at the maximum price
7. Are there any houses re-sold with higher/lower appreciation?
8. Data Cleaning:

Most of the fields found in the King County housing dataset were acceptable for performing analyses. However, while traversing the data few columns require typecasting to achieve requirement. Few columns need to be derived from the existing. Apart from this, there are few duplicate rows based on unique Id needs to be closely scrutinized. Columns are found to have outliers in which some of them are not genuine and might be data entry error. To end with, few (0.0001%) records data is not logically acceptable like year construction is greater than the year it’s sold (exactly 1 year difference). Possibly they could be land lot sold initially and then built in the next year (an un-guided guess).

* **renovated\_flg**: new column added to identify if house is renovated. Refer Table Renovated Flag

Table Renovated Flag

|  |  |
| --- | --- |
| Value | Description |
| 1 | Renovated |
| 0 | Not renovated |

* **house\_age**: new column to determine the age of house when its sold. Basically, the Year of construction and the year of sold columns are used.
* yr\_sold: New Column, Extracted from Date Column calculate: =YEAR([@date])
* mt\_sold: New Column, Extracted from Date Column. calculate: =MONTH([@date])
* Date: Formatted to remove T000000 at end. Verified using the conditional formatting if all the rows end as T000000 calculate =LEFT([@date], LEN([@date])-7)
* seq\_no: new column to identify the house uniquely. Using the autofill function.
* Price : column modified from general to accounting , added $ as prefix
* Bedrooms: Upon analysis found that one record has 33 bedrooms. As per the sqft\_living , the combination of bedroom and bathroom are not in conjunction. Moreover, verifying the address using the Geo spatial sites using the explanatory variables (Latitude and longitude) and observed the neighborhood (8028 Corliss Ave N, Seattle, WA 98103) and assumed the 33 bedrooms a data entry error and corrected it to 3

Results of the analysis:

1. Can we assume the ideal house price based on the bedrooms and bathrooms?

Upon analysis I observed that bedrooms and bathrooms doesn’t decide the price, there are lot more factors like grade, sft\_living, view, condition all together. Refer sheet Initial Analysis in Addepallyassignment1.xlsx

1. Does the view of the house impact the price of the house?

Observation: Prices cannot be predicted based on the view/waterfront, there may be significant consideration and depends on lot more factors along with them. As we see the Max price in each view varies and the view with 3 points has higher price than the one with view 4. Using pivot table concept this analysis is made. Refer sheet pivot table Analysis in Addepallyassignment1.xlsx

1. Renovation of house has any relation to the grade of house.

Observation: Renovation of the house has many other factors need to be considered apart from the Grade of the house and that analysis is considered as out scope. Refer sheet pivot table Analysis in Addepallyassignment1.xlsx

1. Does Condition of the house varies depending on the year of construction, what could be the percentage of houses been sold with poor condition?

Observation: Year of construction has no relation with the condition of the house as there are 90% of the houses constructed in the year 1900 are with good condition. Moreover, irrespective of the year they built only 0.01 % of houses with poor quality very sold and all the other houses sold are with medium to good condition. Refer sheet pivot table Analysis in Addepallyassignment1.xlsx

1. Overall, after how many years the houses are sold to get good returns?

Observation: on an average the houses are sold at 43 years. Refer sheet pivot table Analysis in Addepallyassignment1.xlsx

1. Characteristic of the house sold at the maximum price.

Observation: A very old house after renovation with the highest grade, condition and a medium view house returned a maximum price. Refer to sheet initial analysis in the workbook. Addepallyasignment1.xlsx.

1. Are there any houses re-sold with higher/lower appreciation?

Observation: Yes, there are few houses 0.07% of the dataset are re-sold in year or less than that with slight increase in the price and tiny number of houses are sold at losses. This seems to be like house- flipping was done to gain some profit either by renovating or not

Excel worksheet for reference

