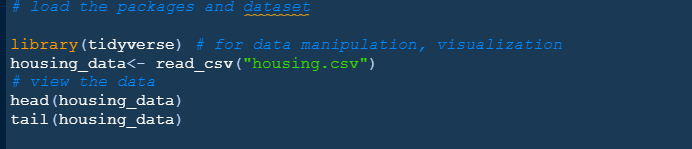
descriptive and exploratory analysis

2022-12-20

**INTRODUCTION**

Our analysis will be conducted using R programming language with the help of rstudio software and other packages, for instance, ggplot2 package that ill be used for plotting. We first start by defining what exploratory and descriptive research means.

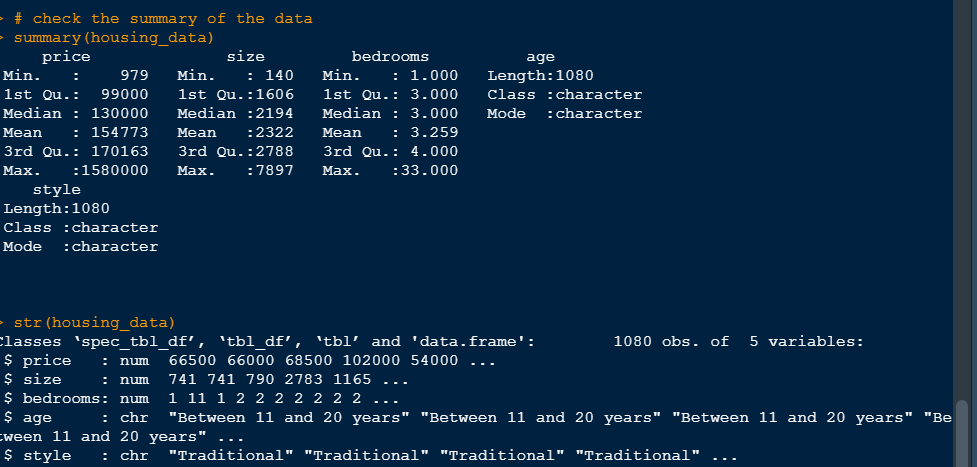
Figure 1: Loading package and dataset



**DESCRIPTIVE ANALYSIS WITH R**

We are going to start with descriptive analysis in R for the provided data above. Using summary(), you can quickly determine the minimum, first quartile, median, mean, third quartile, and maximum for all numerical variables in a dataset. Also str() can give us more structural informations pertaining the dataset.

Figure 2: Codes used to check the summary of the data

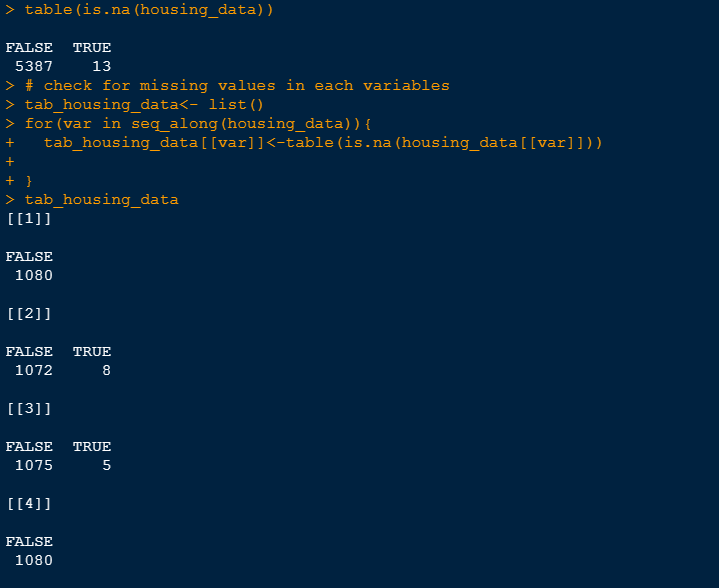


**Handling Missing Values**

Next thing that follows will be to check for missing values in our data. In real-world data, there are several cases where a specific element is missing for a variety of reasons, including corrupt data, information loading issues, or insufficient extraction.

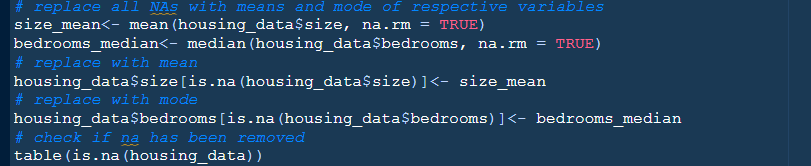
* Substituting Mean, Median, or Mode This tactic can be used on a feature that contains numerical information, such as a person’s age or the cost of a ticket. In order to fill in the missing data, we can compute the feature’s mean, median, or mode. This is an estimate that might introduce variation into the data set. But this approach, which produces better outcomes than removing rows and columns, can negate the loss of the data. We will use this method for analysis.

Figure 3: Checking for missing values



We start by checking whether the missing value present is 13. Using list(),table(), is.na() functions and a for loop we can iterate through the data and create a list that store each table that shows which variables has missing values. We are going to use the mean to replace NA values in size variable and median for bedrooms variables and then check whether NA values are done with.

Figure 4: removing missing values



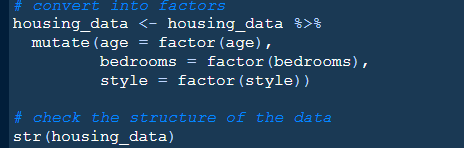
A table with only FALSE variable is returned. Thus all missing values have been delt with accordingly.

**Refactoring Variables in the data**

Some variable such as age, bedrooms, style can be converted to factors. We only make a variable into a factor if one or more of the following conditions are true:

* The variable is categorical in nature since the values of the variable indicate some sort of grouping.
* when character variables have been utilized to define group levels, there are significant memory savings to be achieved. Typically, this is the case.
* There is no better approach to include the variable, which is numerical in nature but extremely non-linear, into a model than to turn it into a factor with one or two significant cut-points selected. We therefore need to convert them into factor and the best way is using the factor() inside the mutate() function from dplyr package as shown below.

Figure 5: Factoring variables

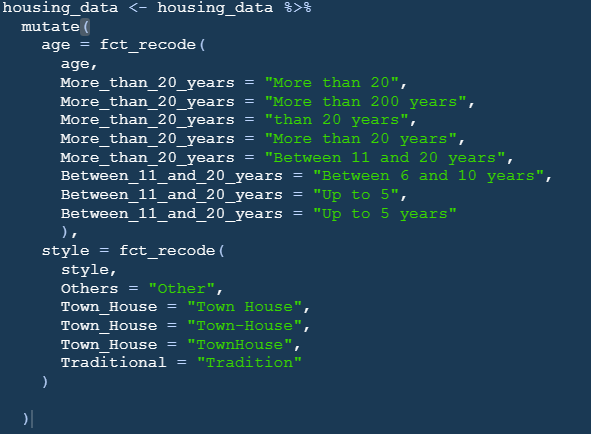


We now see this variables have been converted into factors.

**Cleaning the data**

When we check at age variable, some of this level are outliers for example level like an age bracket of “more than 200 years” is an unrealistic figure therefore should be done with. Also Age like “up to 5 years” is not right since that would mean that a kid can be able to buy a house and that is wrong therefore should be done away with. Age “more than 20” and “more than 20 years” sounds the same so it should also be delt with properly. Age “between 6 and 10 years” can fall under the age bracket “between 11 and 20 years”. Also, the style factor variable has the same features such as having different levels but means the same for instance “Tradition” may be same as “Traditional” but was created maybe due to typing error, “Others” and “Other” are one and the same things and its advisable to clean this factor variable. Therefore using the forcat package from tidyverse meta package, we can refactor this levels.

Figure 6: Changing factors in the dataset

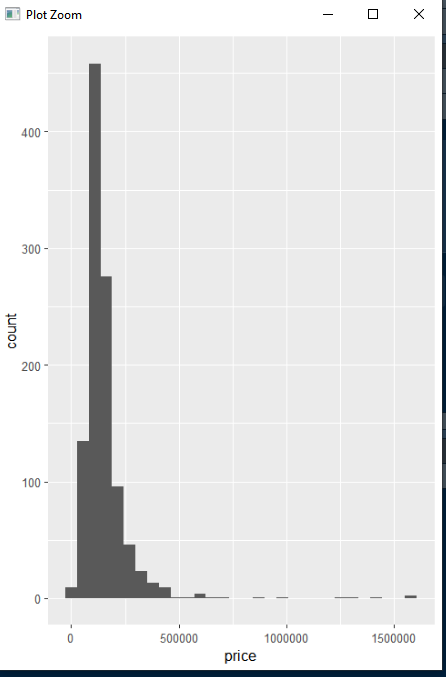


After creating new levels above our data is now clean for further analysis.

**Statistical Properties of Different Variables from the data**

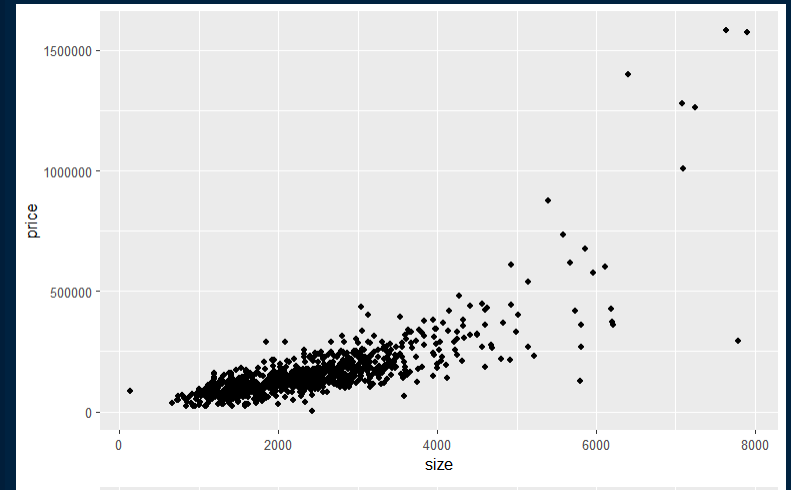
We will start to look at the properties of different variables from the data using some visualizations. Since the variables are only 5 we will look at each one of them looking at some of the relationships of different variables. We can consider looking at the first variable which is price. We can create a histogram that will help us identify how this variable is distributed.

Figure 7: Histogram for House price



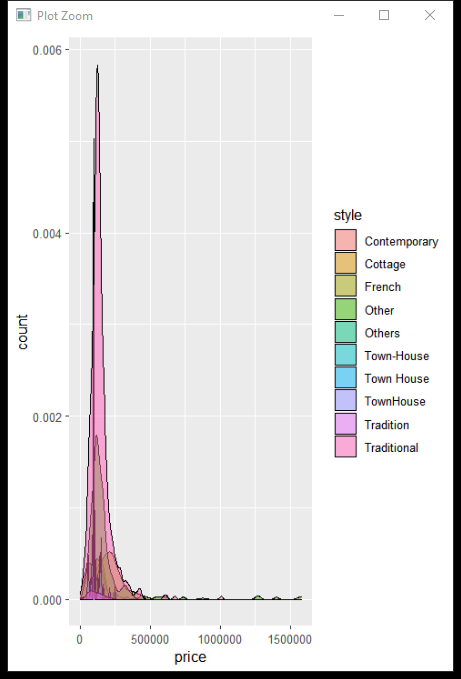
We can see from the histogram that the price variable is skewed to the right with highest count concentrated between 0 and 500000. This gives us a suggestion that the most sold houses have a price ranging between 0 and 500000. The histogram also show us that the highest count of say house sold is about 450 while the least sold houses are below 50 with prices higher than 500000. We can try and see the relationship of the house price together with the houses sizes by creating a scatter plot as shown below.

Figure 8: Scatter plot for the house price versus house sizes



We can see that the price of the house relates to the size of the house in that as the price of the house increases, the size of the house also increases. Although we can also see the houses with sizes less than 4000 the price is less than 500000 thus the concentration lies in this region. Also as the size of the house increase, lower number of houses are sold due to high prices. We can check the relationship between price and style of the houses from our dataset using a density plot as shown below.

Figure 9: density plot for price filled with different house prices

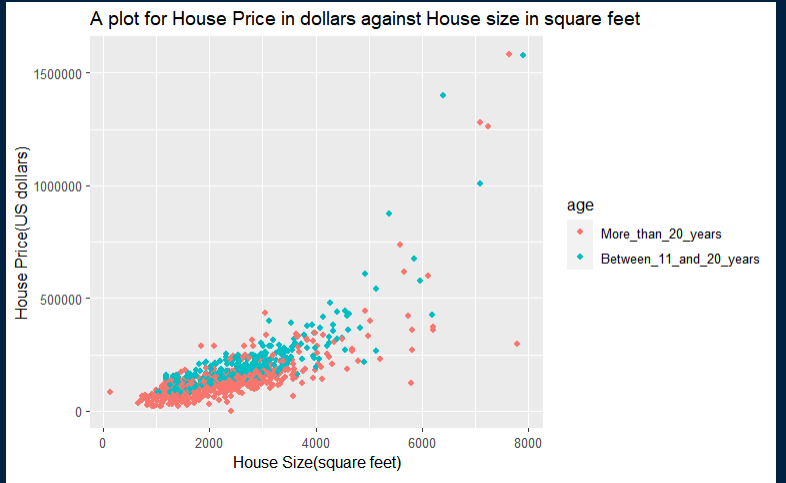


While checking the relationship between the price and the house style we see that the the most common style used to style this houses is the Traditional way while the other styles follows. This style also has a price ranging below 500000.

**Exploratory analysis for housing data**

The main question for this analysis is to predict the house prices as per the given house data. After we have checked our price variable above we would like to further the analysis so that the buyers and and real estate agents can understand more about the house price given other factors. We can start by checking how the house price relates to other variable, starting with a variable called size. A scatter plot is shown below.

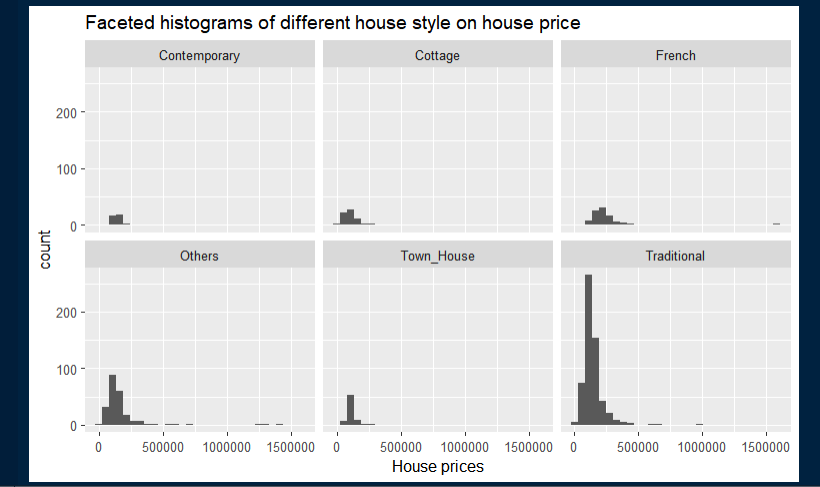
Figure 10: a scatterplot of price and size colored with different ages



By considering that we would like to know which group buys the houses the most, we colored the plot using the age bracket and we clearly see that the house agents target group is those with age more than 20 years.

And by that the age with the highest count is More\_than\_20\_years, ie, age greater than 20 are the most type of people who buys houses the most. We can also create a plot that will show us which house style is bought the most. This can be shown below by creating a faceted histograms plot of house price .

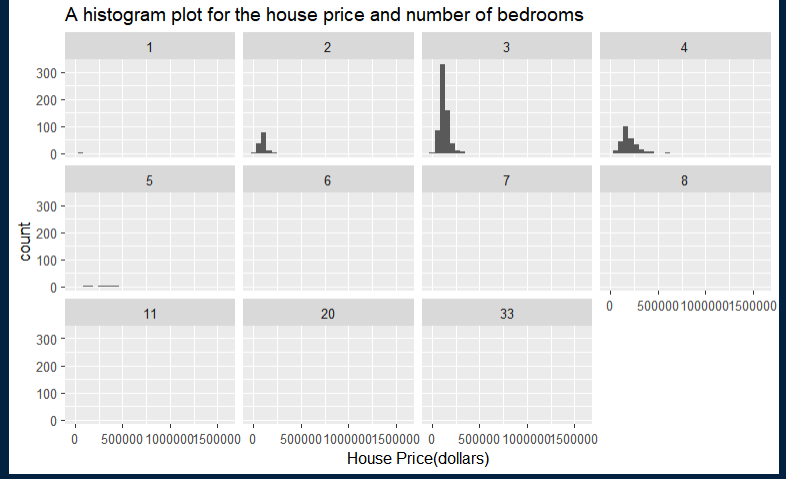
Figure 11: A house price histogram showing different house styles



Most of this house style in our dataset have a count of lower than 100, but we can also see a style that has count more than 250 count is the Traditional style. Although this house has a price of lower than 500000, an agent who would like to build houses for selling will consider the traditional style the most to increase their profits.

Lastly, looking at the number of bedrooms, we have to come up with a visualizations that will show us what number of bedrooms sold well and was well considered.

Figure 12: A histogram showing different numbers of bedrooms



We can see that the most house bought had 3 number of bedrooms followed by 4 then 2. Therefore its well clear to the house agents that the house that is preferred to sell should have 3 bedroom for gain of maximum profits.

# CONCLUSION

We can coclude that:

* an agent can make more profits if he/she invest his money and time on the Traditional house style, for better profits.
* If one wants to buy a house he/she might consider house that has Traditional style in it.
* We also saw through the analysis that most people who buy houses have age above 20 years.
* We also saw in the above analysis that as the size of the house increases, the price also increases.
* Also, if an agent want to increase profits earned in this sector, he should sell houses with 3 bedrooms since most people are more likely to buy this house.

# Reference

[Surbhi S](https://keydifferences.com/author/surbhi). 2017, *Difference between Exploratory and descriptive research.*   Available from <[Difference Between Exploratory and Descriptive Research (with Comparison Chart) - Key Differences](https://keydifferences.com/difference-between-exploratory-and-descriptive-research.html#:~:text=Exploratory%20research%20is%20one%20which%20aims%20at%20providing,out%20research%20activities%20in%20different%20fields%20of%20study.)> [September 2, 2017].

Soetewey A. 2020. *Descriptive statistics in R.* available from <https://statsandr.com/blog/descriptive-statistics-in-r/>. [January 01, 2020].