**COMPARING HOUSE PRICES AND CRIME RATES IN NY AND FL LOCATION: AN ANALYSIS**

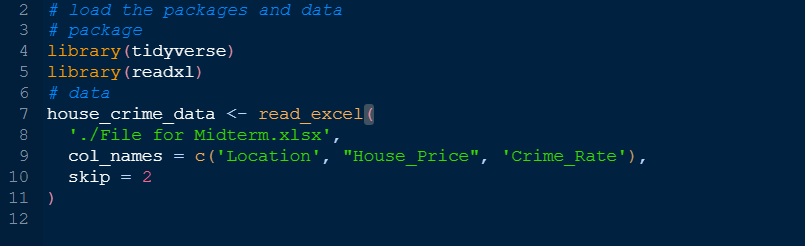
**INTRODUCTION**

It can be both exciting and difficult to move to a new location, particularly if you want to make sure that you and your family will be safe and secure. Suzie, for instance, must choose between moving to FL and NY location as per the given data. Although she needs to take affordability into account, she wishes to live in a neighborhood with lower crime rates. Her agent has given her information on crime statistics and home prices for properties that meet her criteria so that she can make an educated choice. This report aims to analyze the information given by the agent and offer insights into how house prices and crime rates relate to one another in various places. Suzie will be able to make a choice that satisfies her requirements and tastes thanks to the report's explanation of the compromises between the two variables.

**LOADING THE DATA AND REQUIRED PACKAGES**

The Excel file containing the three variables -- Location, House\_Price, and Crime\_Rate -- will be imported into the analysis software in the section titled loading data. After reading the data, the program creates a dataset with the variable columns that correlate to it. Utilizing different statistical techniques and machine learning algorithms, the dataset can then be explored, visualized, and analyzed in order to gain insights and make defensible decisions based on the data given by the variables. The "read\_excel" function from the "readxl" package would be used in the R program's loading data subsection to read the Excel file. To manipulate and prepare the data for further research, use the "tidyverse" package.

The first step is to import the data and required packages using the following command:

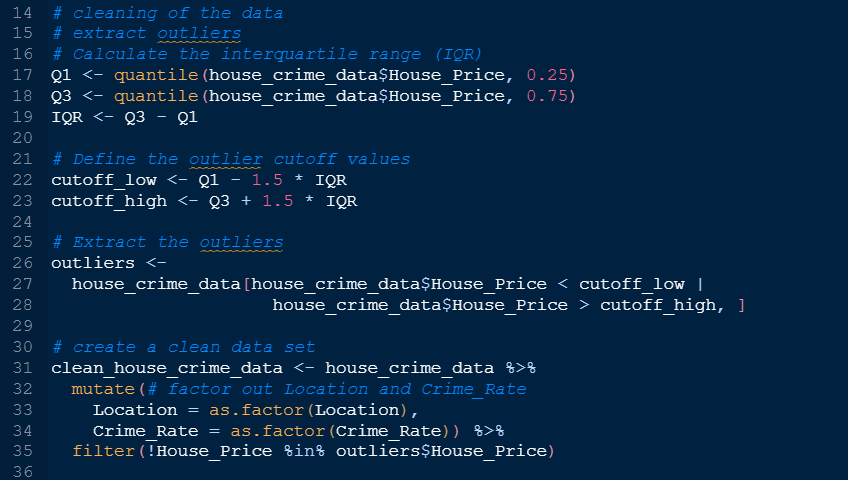


**CLEANING OF THE DATA**

After loading the data, we need to clean it for better results. In order to ensure that the data is accurate, consistent, and prepared for further analysis, cleaning the data is an essential stage in the data analysis process. Various R functions and methods can be used to manipulate and transform the data to make it more useful and meaningful. This is covered in this subtopic.

Overall, guaranteeing the accuracy and validity of any following analysis begins with cleaning the data, which can be accomplished by utilizing the proper R functions and techniques.

Cleaning of data is shown in the steps below.



We are now going to use the clean dataset created above to make analysis.

**DATA ANALYSIS**

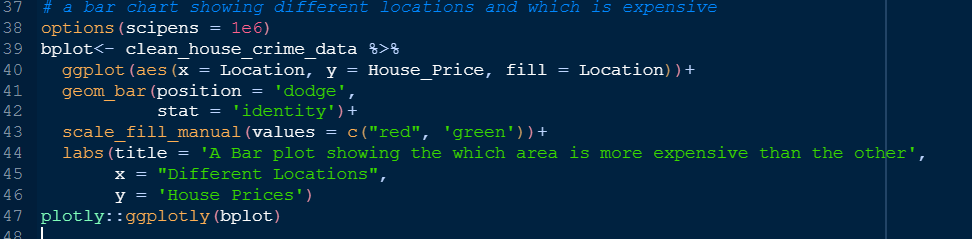
Suzie had other consideration in mind before she moved to either of the two places. First she would like a place that was low in crime and other factors were:

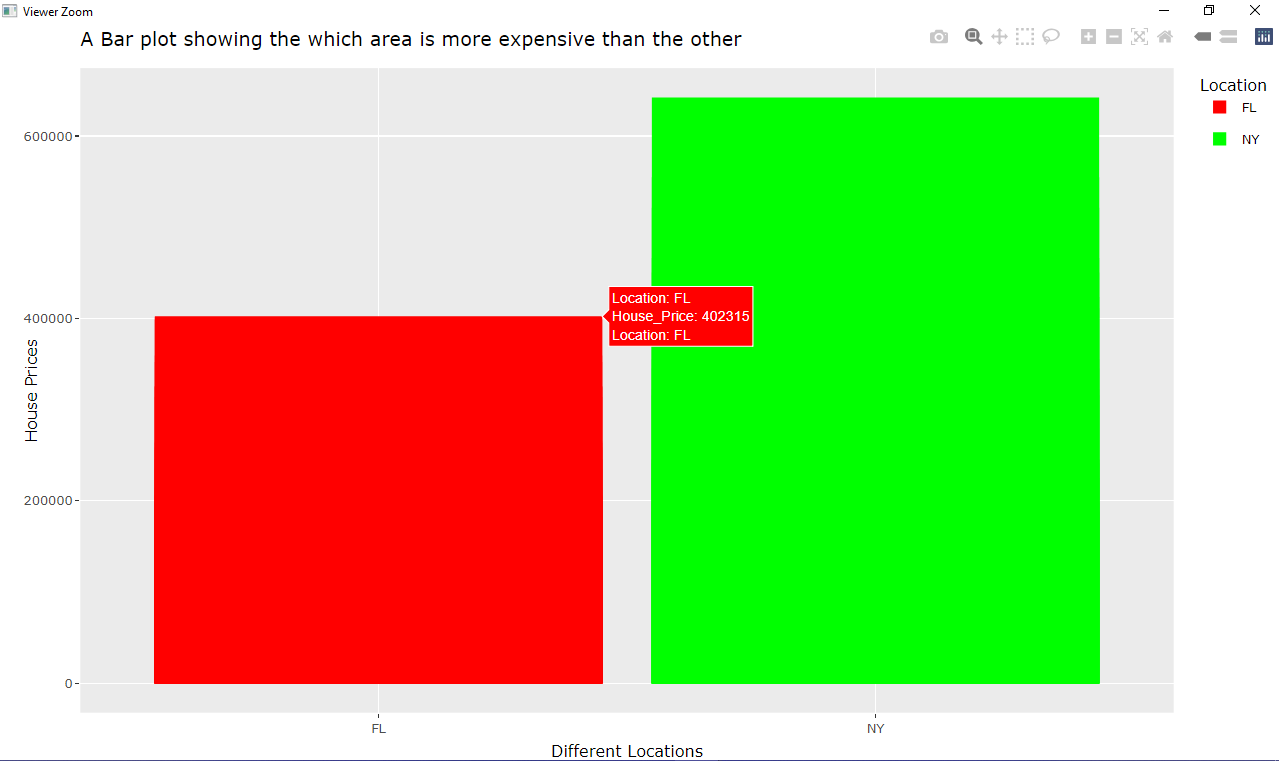
1. Is it more expensive to reside in NY or FL, or is one more affordable?
2. Which state has a higher crime incidence, FL or NY?
3. Is crime more prevalent or less prevalent in regions with more expensive or less expensive housing?

These questions will form the basis of our analysis. We will start with the first question and determine which area was more expensive to live in or which was affordable to reside.

1. **Is it more expensive to reside in NY or FL, or is one more affordable?**

We are going to use a visualization so that we can understand the question asked. The best plot will be a bar chart which will show the size of each location compared to the house prices in each location. A screenshot below shows how this was done.

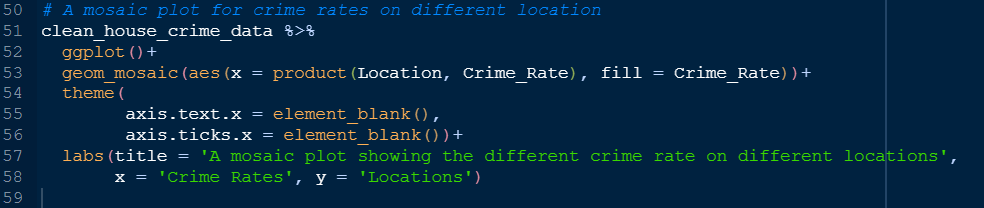


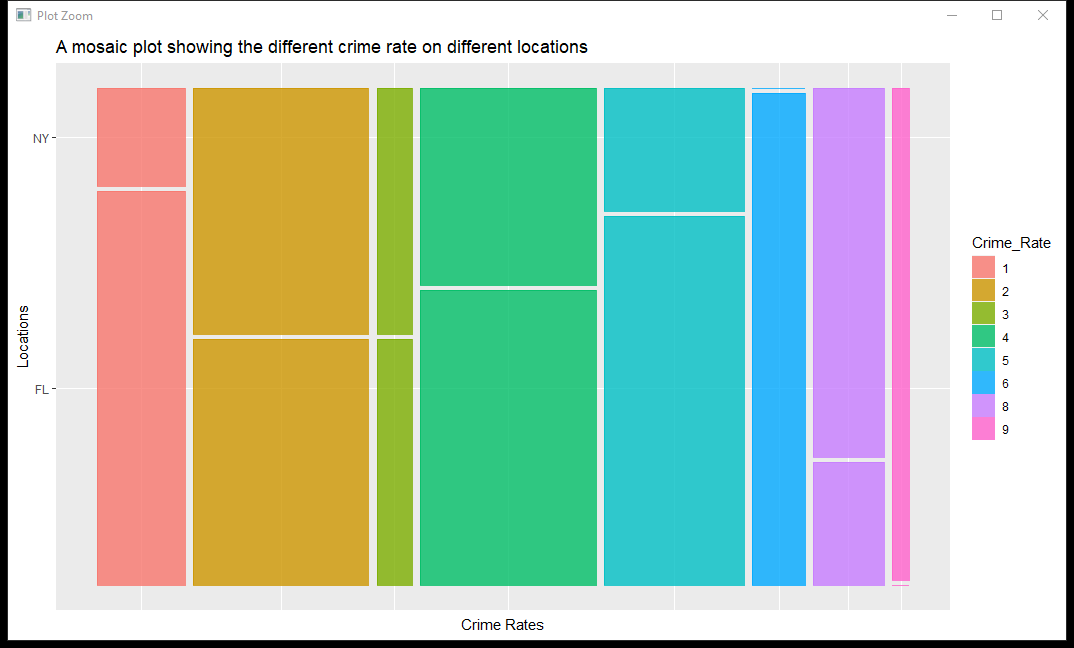


In the bar plot above we have created, we see the locations are plotted on the x-axis while the house prices are plotted on the y-axis. As the house prices increases so does the bar for the NY location. Since the bar representing NY is bigger than the bar representing FL in a bar chart where we are comparing house data from two different places (i.e., NY and FL), it denotes that the average house price in NY is greater than the average house price in FL. This may give us a suggestion that as the house price gets high so does the cost of living in this area thus we can conclude and say that NY location is more expensive to live in compared to FL location.

1. **Which state has a high crime incidence, FL and NY?**

For this analysis we are supposed to create a plot that will help us identify the relationship between the crime rates on the different locations. A mosaic plot in R is a graphical representation that displays the proportion or relative frequency of two or more categorical factors. The associations between categorical variables can be seen using mosaic plots, which can also reveal patterns or relationships that may not be clear from other kinds of plots. Each rectangle in a mosaic plot depicts a cell in a contingency table, and the rectangle's size is proportional to the number of cells or their frequency. The plot is created by stacking the columns and the squares in order to create the columns. The geom\_mosaic() function from the ggmosaic package in R is used to make mosaic plots and provides for customization of the plot's appearance and labeling of the axes and margins. This is shown below how it was done.

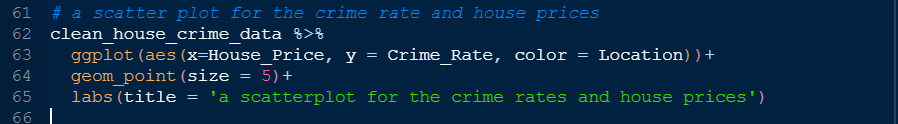


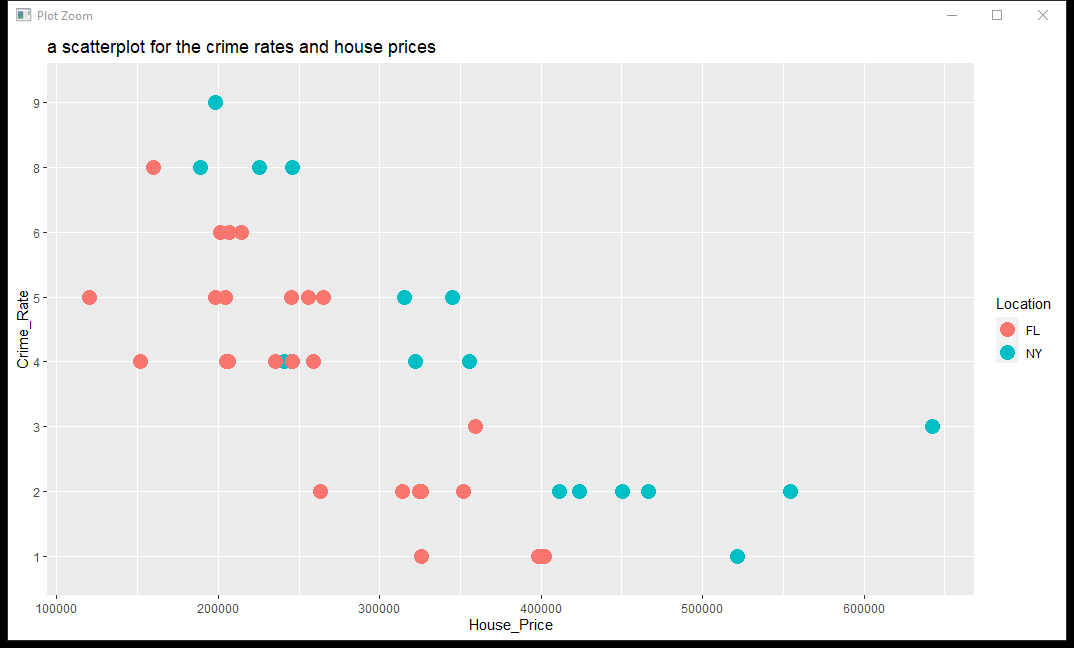


The plot above is a mosaic plot made using R. we see clearly the different levels of crime rates starting from 1 through 9 where 1 is ranked as lowest crime rate and 9 being the highest. Starting with 1, we can see that the level is divided into 2 location, on the top the NY occupies this level and on the bottom the FL occupies this area. Since 1 is a low crime rate we can see that low crime rates are high in FL location by looking at the area covered by the FL location. The NY has low crime rate in this level. Considering the largest crime rate, 9, we see that the area representing the NY region is high compared to FL. This can be due to NY might have high cost of living thus most crime occurs in this area.

1. **Is crime more prevalent or less prevalent in regions with more expensive or less** **expensive housing?**

The most effective form of plot for examining the correlation between crime rates and home prices is a scatterplot. On the y-axis, we can plot the price of homes, and on the x-axis, we can plot the crime rate. A region or place is represented by each point on the scatterplot. We can determine any patterns or trends in the data using the scatterplot, such as whether areas with higher house values typically have lower crime rates or the opposite is true. To see if these elements influence the correlation between crime rates and home prices, we can also use color or shape to separate regions. All things considered, a scatterplot is a flexible and useful tool for examining the correlation between two continuous variables and can assist in determining whether crime is more or less common in areas with more or less costly housing. This is shown below using a screenshot.





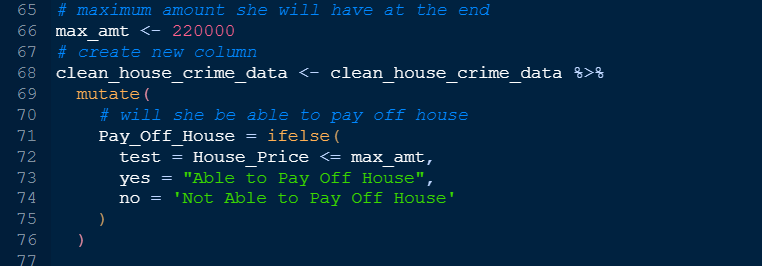
In the above plot we see that as the house price increases the crime rates decreases. The points in the far end have crime rate of below 3 when the house price is above 500000. We also see that as the house prices reduces, the crime rate levels increase. When the house price is below 400000 the crime rate increases from 1 through 8. Therefore we can conclude and say that as the crime rate transition from 1 to 9, that is, as the crime rate increases, the house price reduces. Thus, crime rate is high in area with low house price and low in houses with high prices. Although NY has high house prices it also have high crime rate recorded, while FL has low house prices with crime rates spread between different levels.

1. **Where would I move to if I was Suzie?**

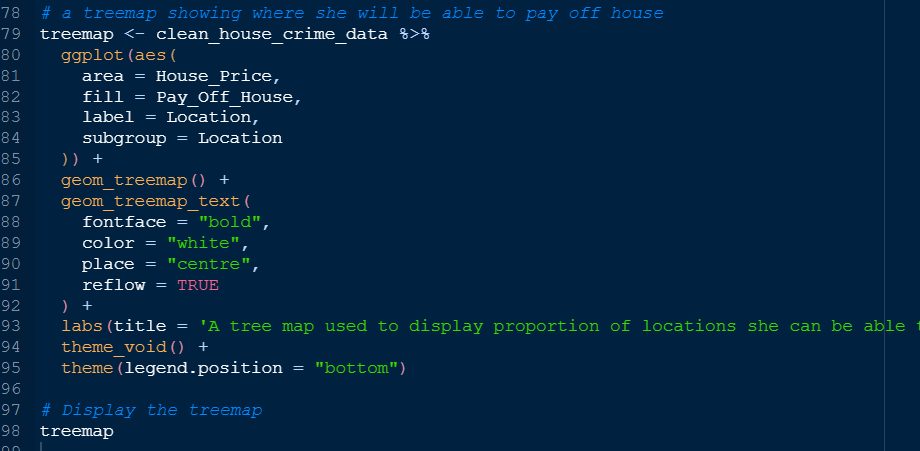
Since Suzie would like an area with low crime rates then best suggestion will be to move to FL since FL has low crime rate.

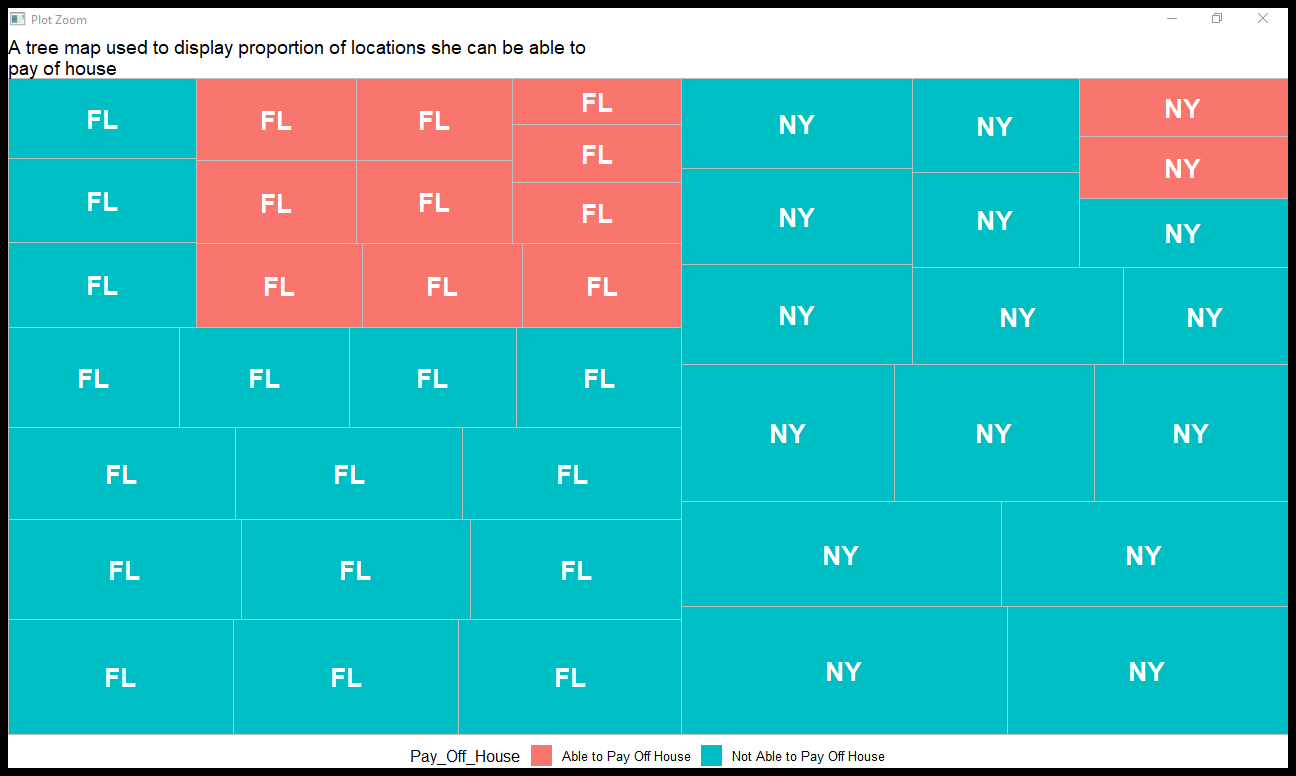
1. Additional information was provided and the following was required.
2. What location will she be able to pay off considering the average house price set aside and the income she will earn?
3. Where should she move to and why?
4. **What location will she be able to pay off considering the average house price set aside** **and the income she will earn?**

The amount she will have if she moves to NY is 220000 while if she moves to FL will be 175000 thus from the data therefore we need to create a new column with 2 levels which are ‘Able to pay off house’ and ‘not able to pay off house’. This column will be created such that the data that representing ‘able to pay off the house’ level should be less than the maximum amount which is 220000 and the rest of data where amounts of house is greater than this will be set to ‘not able to pay off house’ level. This is shown in the codes below.



While the screenshot showing the treemap codes is displayed below.





As in the above diagram we see that the FL area has a high proportionality for her to live here according to the stated conditions. Therefore, we conclude and say that she has a high probability of being able to pay off houses in FL than NY.

1. **Where should she move and why?**

Suzie will be able to move to FL due to: FL has low crime rates, house price in this area are affordable and due to the amount set aside and the income she will earn she will be able to pay off house in this area.