**ASSIGNMENT: ANALYSIS ON HOUSE PRICING AND CRIME RATES**

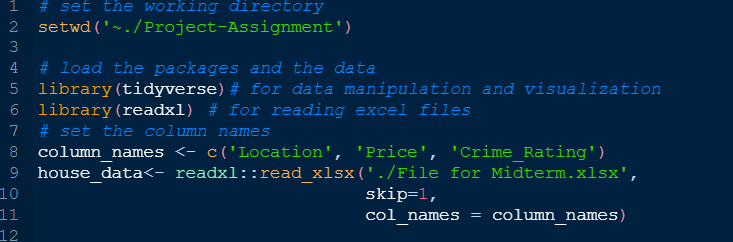
**INTRODUCTION**

In our assignment, we are supposed to analysis some data for Suzie so that she can have a clear understanding of the house prices and the crime rates in the location she wants to move to, therefore we need to answer some question using the data. The data is stored in an excel extension file which should be read in R using the available functions. This data is divided into three columns: location, home price, and criminal severity. A component with two levels, NY and FL, which stand in for the two states being considered, makes up the location variable. With NY typically having higher prices than FL, the house price variable gives information on how much houses cost in each area. Each location's level of crime is indicated by the crime rating variable, with NY typically having a higher incidence of crime than FL. It is feasible to learn more about the connection between place, home prices, and crime rates by looking at these factors. Suzie, who is looking for the best place to relocate in light of some other factors that will be examined below, can benefit from understanding these factors.

**LOADING AND CLEANING THE DATA**

Before we start any analysis, the data should be loaded and cleaned. To make sure a dataset is accurate, full, and consistent, cleaning the data entails locating and fixing errors, inconsistencies, and inaccuracies. This procedure is essential for removing possible bias or error sources, enhancing the validity and reliability of the findings, and boosting confidence in the interpretations of the data. Data cleaning is an essential phase in the information analysis procedure that can have a big effect on the accuracy and value of the outcomes.

Therefore in our case, since the data is of excel file extension, we will use the read\_xlsx() function from the readxl package to load the data to the workspace. These packages that we will use will also be loaded. Loading of data can be shown below with a screenshot.

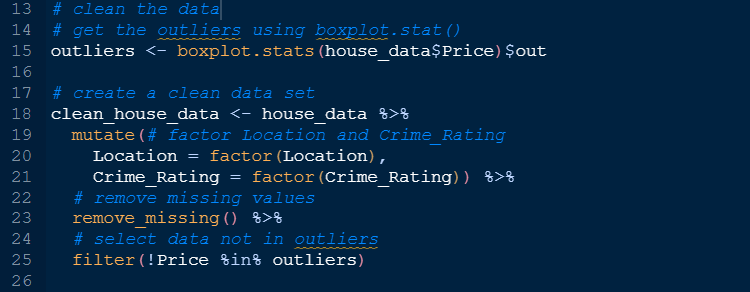


From above screenshot, we first started by setting the working directory using the setwd() which help make data analysis initiatives more effective, accurate, and reproducible. Then using the library() function we loaded the packages that will be used in this analysis. In line 8, the column names were set and then an object called ‘house\_data’ was created.

Next, we need to clean the data and make sure;

1. There is no missing data or row with missing values,
2. The data has no outliers, that is, no variables have extreme values,
3. Each variable has its data type, that is, if a variable is categorical then should be converted to factor.

This was done in R and a screenshot below show each step.



We first started by selecting the outliers using the boxplot.stat function and they will be stored in ‘out’ value returned by this function and store then in ‘outliers’ object. We then using the tidyverse data manipulation technique, we use mutate function to convert the Location and Crime\_Rating into factors, remove\_missing function which removes rows that has NA in it and finally using filter function to select the data that is not in outliers object created above. Now the data is clean for analysis.

**ANALYSIS AND DATA VISUALIZATION**

This analysis will be conducted for Suzie since she is trying to find the best place to move in with some few considerations. These factors to consider are the one the analysis will be conducted on. Suzie wants to live in a location with low crime rate. Other factors she considered were:

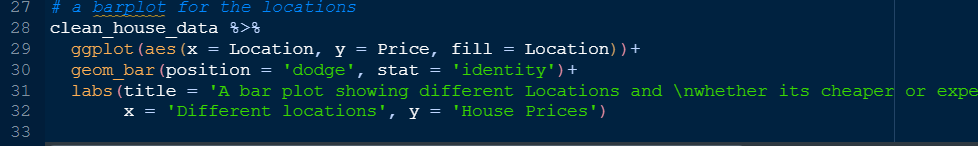
1. Can you compare the cost of living in FL and NY and determine which location is cheaper or more expensive?
2. Can you compare the crime rates of NY and FL and determine which location has a higher or lower crime rate?
3. Is there a correlation between high housing prices and high crime rates in a particular area?

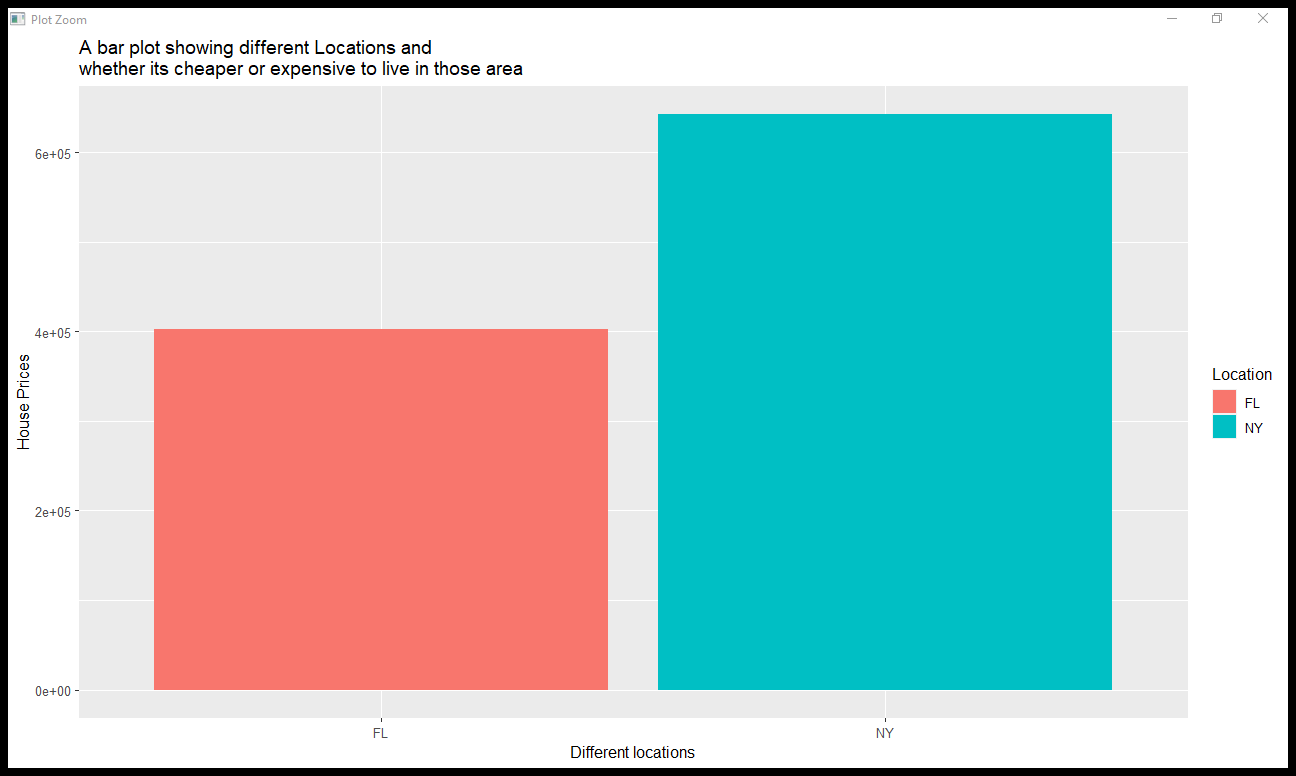
We first start by making an analysis to these question that Suzie puts into consideration.

1. **Can you compare the cost of living in FL and NY and determine which location is cheaper or more expensive?**

The best way to analyze this question will be to use a proper visualization and the best plotting method will be a bar plot. The outcomes of categorical factors or discrete data are frequently displayed and compared using bar plots, also referred to as bar charts. When comparing the relative sizes of various categories or displaying the occurrence distribution of a variable, bar plots are especially helpful. The Crime\_Rating will be the category variable in this instance.

Using R, we can create a plot and display it for analysis and this will be shown below.

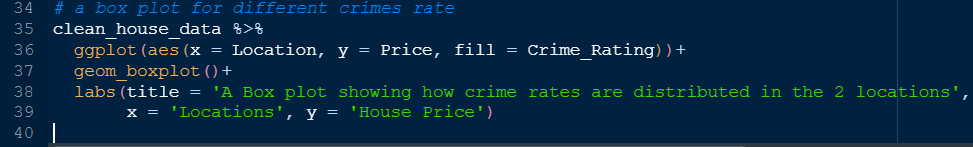


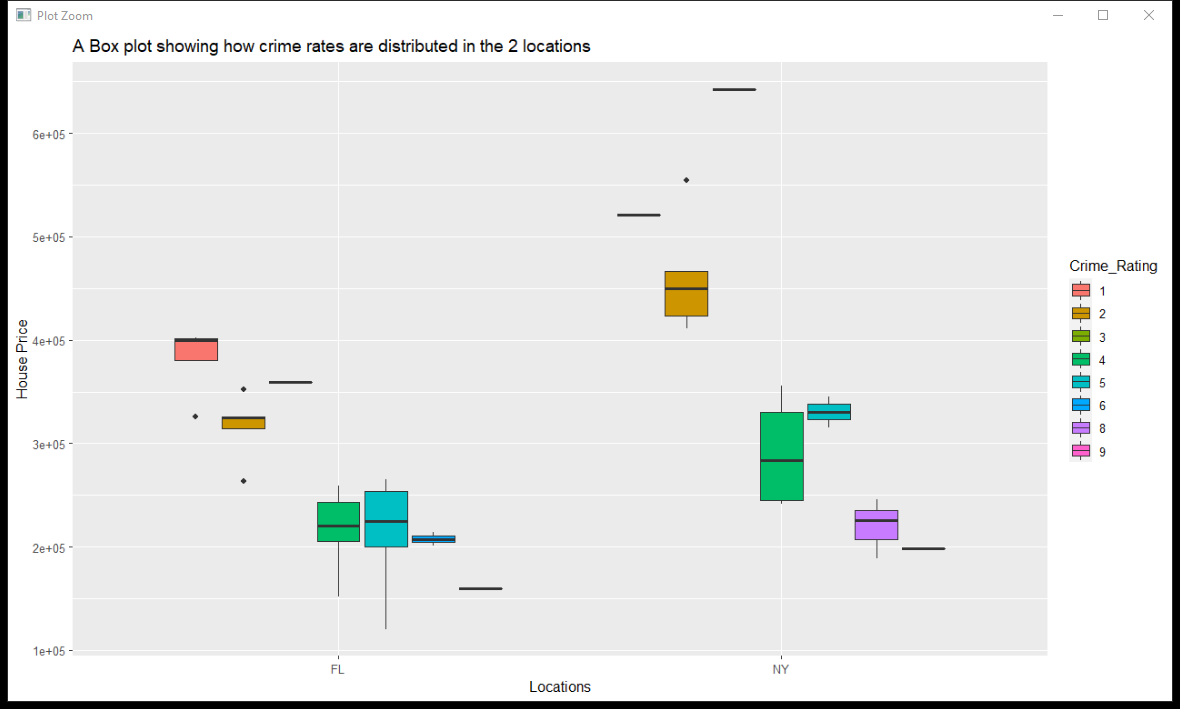


Using the geom\_bar function in the code shown above, we created a bar plot that shows which location is expensive or cheaper to live in. On the x-axis we have used the Location variable as our dependent variable and we see there are only 2 locations present in the data by looking at the legend in the plot while the price of the houses are used as independent variable with values ranging from 0 to 600000. The FL location is the one on the left which has a low bar. represent a smaller value or a lower frequency compared to the other bars in the plot. This means that the FL location is cheaper to live in with houses prices lower than 400000. The NY location has a larger bar compared to the one on the left. In a bar plot, a bigger bar indicates a greater frequency or higher value in comparison to the remainder of the bars in the plot. In this case, the bar signifies that the place is expensive to live in due to the price of houses present in this area. The house prices are high with values as high as 600000 and this says that the cost of living in this are is expensive.

1. **Can you compare the crime rates of NY and FL and determine which location has a** **higher or lower crime rate?**

Next, we can consider another factor that Suzie put into consideration before selecting an area to live in. This factor is the crime rate and determining which location has high or low crime rate. Codes displayed below helped us come up with a box plot for this analysis.

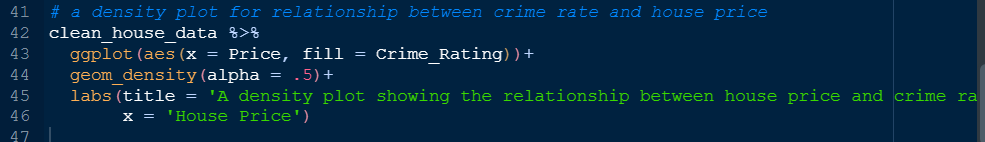


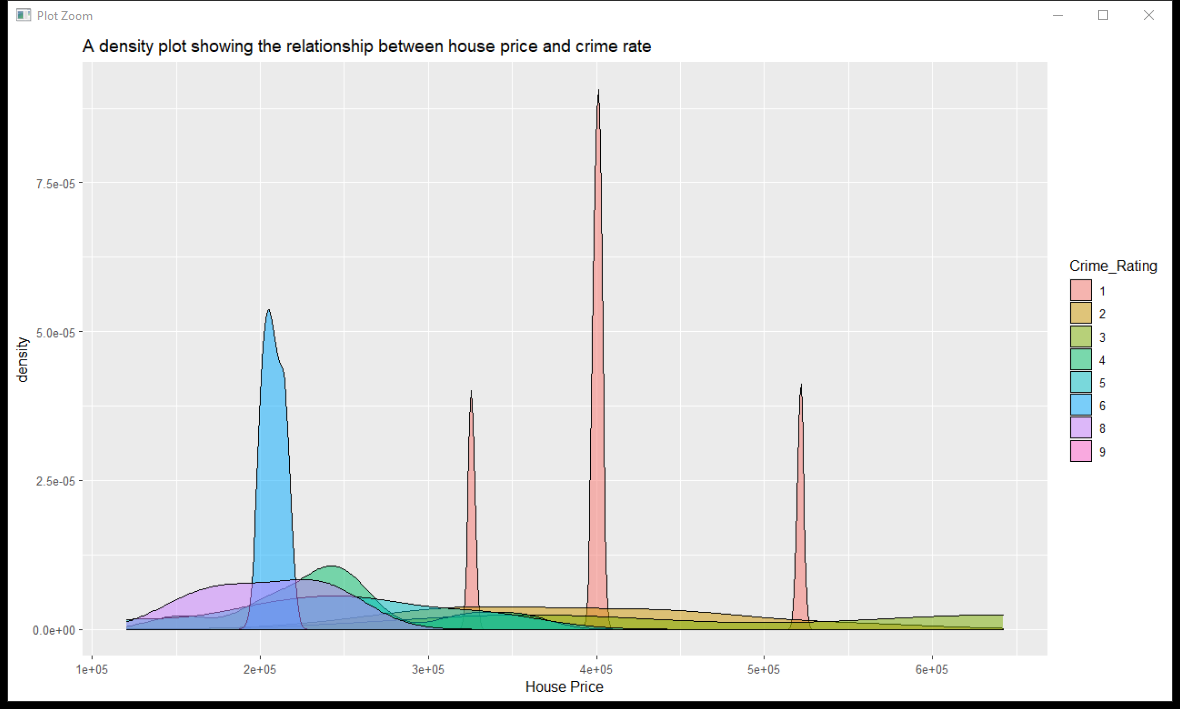


A box plot, also called a box and whisker plot, is a visual depiction of a group of data that highlights how the data are distributed. The interquartile range (IQR), or box in the plot, is the range of numbers that contains the middle 50% of the data. The median value, or the middle value when the numbers are arranged in order, is represented by the line in the box. A box plot's dimensions can reveal some details about the distribution of the data. Compared to a smaller box plot, a larger box plot typically denotes a broader range of values and more variation in the data. For instance, as the house prices increases in FL, the level of crime rate in the FL reduces from 9 through 1 given that 9 is the highest level and 1 the lowest. Similarly, this is seen in the NY location. But in NY location there is a higher distribution of crime level 8 compared to the FL location. This high crime rate in NY can be concluded to be as the house prices reduces then the crime rate gets high.

1. **Is there a correlation between high housing prices and high crime rates in a particular area?**

We will use the variables given to come up with a relationship between the house price and crime rate. A density plot will be best at giving this analysis.





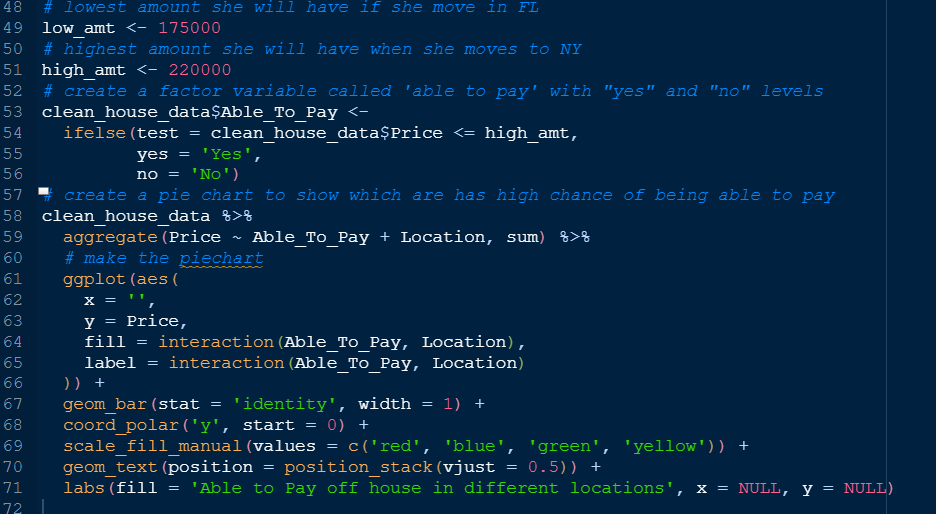
In this plot, we see a relationship between two variables, prices of the houses and the crime rate. The crime rate has 9 levels with 1 being the lowest crime rate and 9 the highest crime rate while the house price ranges between 100000 to 600000. We see when the house price is low, less than 300000, the crime rate is of levels 6 to 8 while as the prices increases the crime rate reduces with 1 having high density meaning the crime rate is low. Other lower levels such as 2 and 3 spread across the graph as the price of houses increases. We can conclude and say that area with low house prices have a tendency of having high crime rates compared to areas with high house prices.

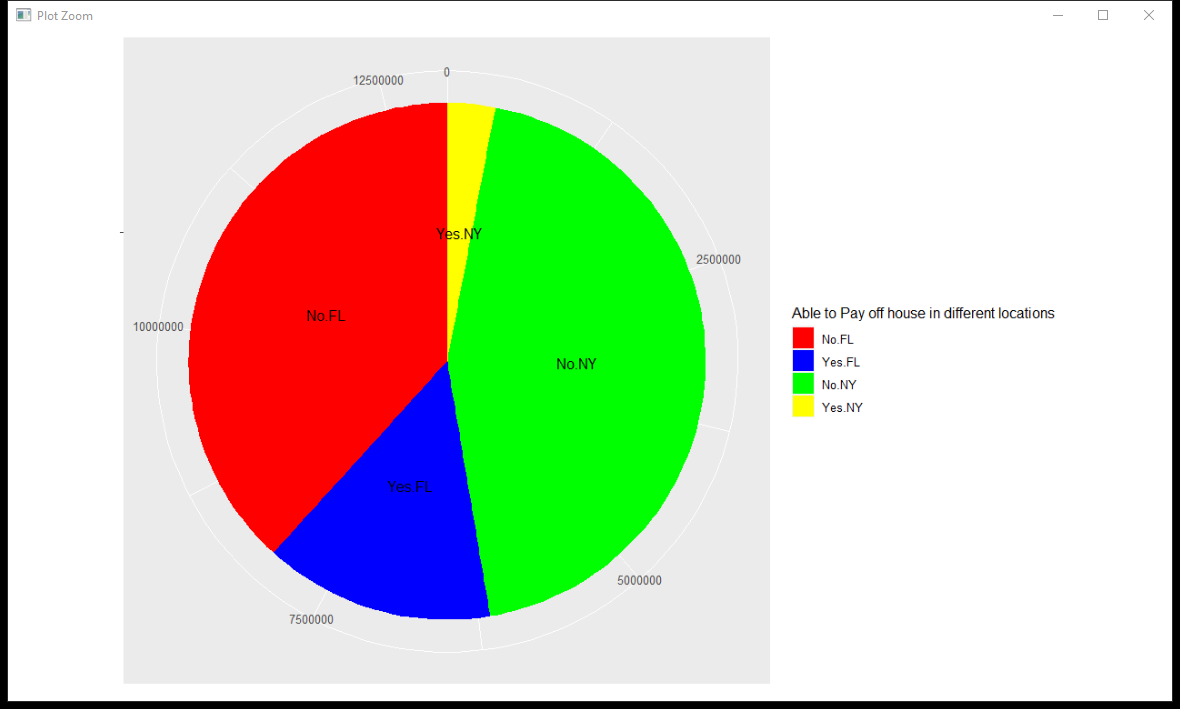
1. **Based on the analysis, where should Suzie move to?**

Since Suzie wants a place with low crime rate, the suggestion would be to move to FL since FL has low crime rates as per the analysis above and houses are cheap in this area.

1. Some additional information was given out by Suzie and the following is required.
2. What location will she be able to pay off the house considering the average house price set aside and the income she will receive?
3. Where should she move and why?
4. **Which location will she be able to pay off the house given the average house price and income she will receive?**

Given that Suzie set aside $100000 for the house she wants to buy and that the income she will earn if she moves to NY is 120000 per year and if she move to FL she will earn an income of $75000 per year. We need to come up with a visualization that will help us know if she will be able to pay of the house. Screenshots of the code are displayed and explained below.





We first created a new factor variable called ‘Able\_To\_Pay’ with 2 factors levels, ‘Yes’ and “No” where ‘Yes’ represents that she will be able to pay the house off while ‘No’ represent that she will not be able to. After that we created a pie chart showing where she is going to be able to pay the house. We see that she have a higher percentage of being able to pay off house in FL than in NY location. Although we have high ratio of the pie that shows that she will not be able to pay off house in both area. This means that the house prices are above the bracket of the income plus the average salary thus the total value is higher than what she is setting aside and the income she will receive.

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

After conducting the analysis above we can conclude and say that Suzie should move to FL location due to:

1. The houses price in this area is cheap and affordable.
2. The crime rate is low compared to NY place which has high house prices
3. According to the amount of money she sets aside for the house and the income she will receive; she will afford houses in FL region as shown in the above analyses.