**5. Marital Status,**

Marital Status of the employees can be a determining factor regarding a lot of aspects in this HR analytics dataset. At first glance, u can notice from the columns in the dataset that it could be related to a lot of the columns (areas) that are mentioned.

Therefore, we will try to analyze these data so we can get more insights, information, identify patterns etc.

**Data Visualization**

We’ve used several types of plots in trying to gain insights from the data analyzed related to marital status. Although since marital status is a categorical data variable, we’ve used more types of plots like boxplots and bar-plots as they are the ones that are mostly used to visualize categorical data type variables.

Although we can also convert these categorical variables to numerical data type variables by encoding them into a factor variable and then use mapping to convert them into a numerical variable. The following code can be used for this process but, we haven’t used it in data visualization as it wasn’t necessary.

A computer screen with text

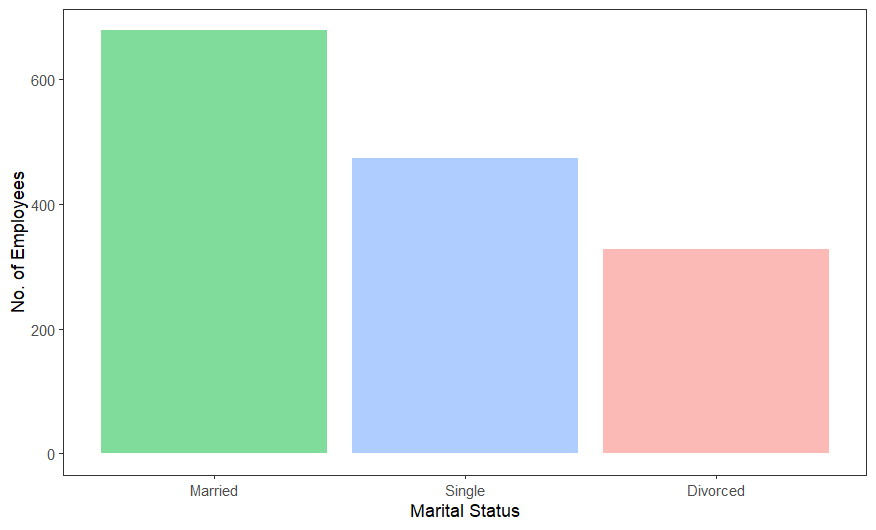
Description automatically generated

Now we are going to look to gain some insights from the plots we’ve generated using R. But first we must install the necessary libraries and load the dataset we are using into a data frame in a R script file.

A black screen with green text

Description automatically generated

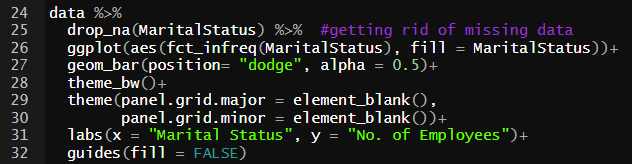
We’ve used the above code for that process and to view and analyze the structure of the dataset we will be using.



The above bar-plot was created to get an idea of the number of employees based on their marital status.

From this we can see that most of the employees are married (more than 600). There’re about 500 employees that are still single. And there’re around 300 employees that are divorced.

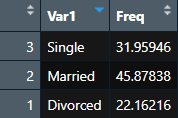
The following code was used to generate the above plot.



And we’ve also used the following code to create a table of the percentages of the marital status of the employees.



The calculated percentages of the marital status of the employees are,



**Marital Status related with Age of the Employees,**

A graph of different colored boxes

Description automatically generated

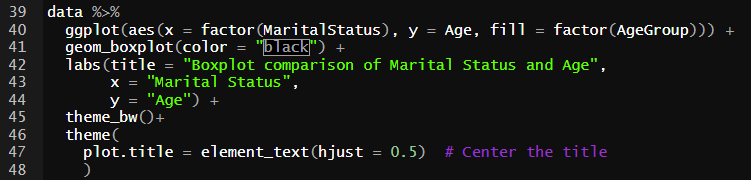
We can use this boxplot to say a lot of things about employees’ ages and marital status.

From this plot as we have factored age groups, we can clearly see that there are a smaller number of people who are divorced between ages 18-25. And there’s a larger number of people who are divorced from ages 46 to 55. We can also see there are some existing outliers around the age of 60 which means there are a few people who are still divorced around that age.

Out of the married employees, we can see there’s only a few numbers of employees who are married between the ages of 18-25. And with the rest of the age groups, it’s kind of divided evenly.

As for the single employees, we can see there are lots of employees who are still single around the ages 18-25; whereas there were a lesser number of people who were married or divorced in that age group. Which only makes sense as that is the youngest age group and it’s more likely that they are mostly single and only a few employees are divorced because that is like the youngest age one would get married. And we can also see how there’s only a few numbers of employees who are still single within ages of 46-55, and even more less in 55+ aged employees; it is kind of self-explanatory as it is very unlikely that people are still single around older ages (not that it can’t be the case but by nature it is not how it is).

We’ve used the following code to generate the above plot,



**Marital Status based on Gender of the Employees,**

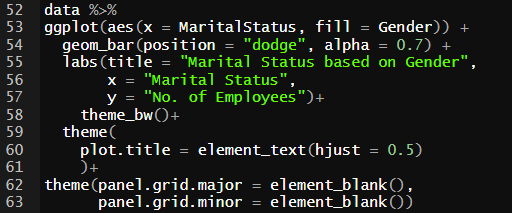
A graph of different colored bars

Description automatically generated

With data visualization using R, we can easily get an understanding of the distribution of marital status of the employees according to their respective genders from the above bar graph.

We can see most of the employees that are married are male employees. Or with any category the number of male employees outnumber the female employees; it would make sense as there’s an exceedingly large number of male employees in this HR analytics dataset. However, we can still make observations such as there’s a 1:2 ratio between male and female employees respectively who are divorced. We can see there’s not that much of a difference in numbers between still single male or female employees.

The following code was used to generate the above bar-plot.



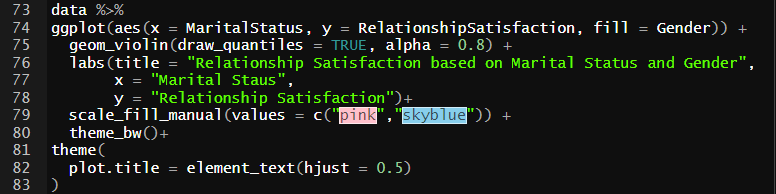
Here are a few other plots we’ve generated to get an understanding of **Relationship Satisfaction**.

A graph of different colors of men and women

Description automatically generated

This graph shows us the density of the number of employees with their relative Relationship Satisfaction level with gender as a factor. We can see that there’re a lot of divorced female employees with a Relationship Satisfaction level of 3. There’re only a few observations like that you can make from this graph.

We’ve used the following code to generate the above violin plot.



The following graph shows us the distribution the of number of employees in different age groups according to their relationship satisfaction level with gender also as a factor.

A graph of different age groups

Description automatically generated

The following code was used to generate the above bar graphs.

A computer screen shot of text

Description automatically generated

The following plot is a density plot between the Distance to Work from Home of the employees and their Marital Status.

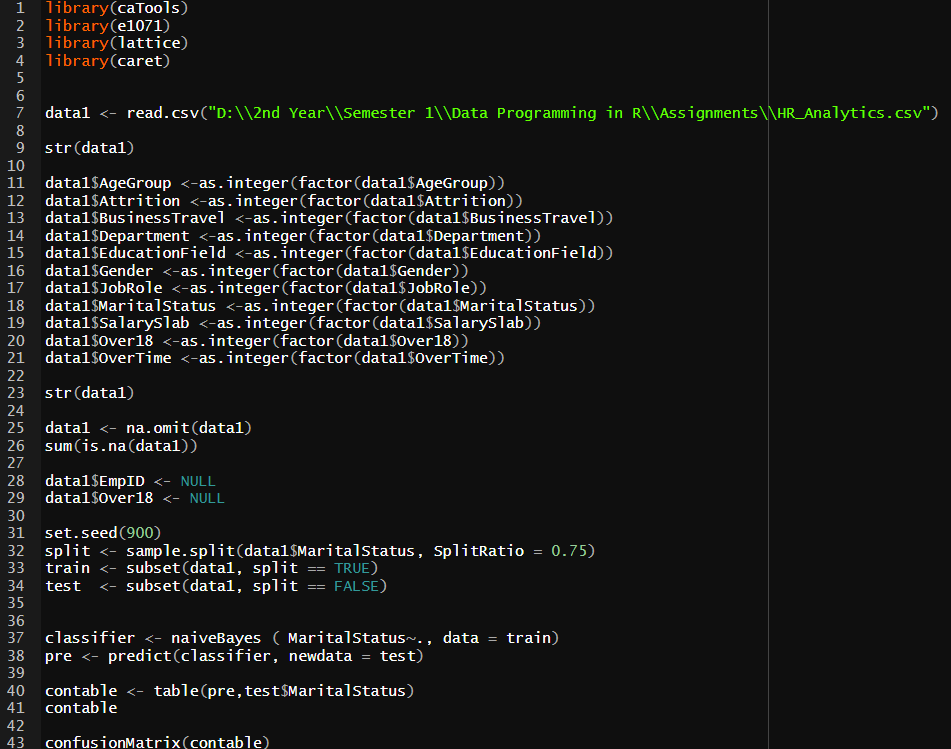
A graph of different colored lines

Description automatically generated

**Data Analytics through models**

**Naïve Bayes model,**

Naïve bayes is a probabilistic machine learning algorithm based on Bayes’ theorem. It assumes that features are conditionally independent given the class label, which simplifies computations. So, we are going to use this model to calculate the predictability of the variable – Marital Status.

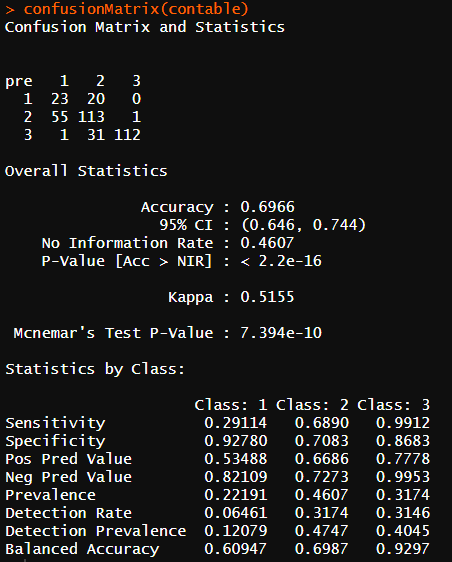


First, we install the necessary libraries and load the dataset we are going to use. Then we identify the structure of the dataset and convert all the categorical variables into factorial variables and then convert them to integer variables.

Then we get rid of missing values in the dataset and check to see if there are any left.

Then we get rid of the columns which shouldn’t be a part of the naïve bayes model.

We get the following output from the above code.



From this, we can say that we can predict the outcome of the variable Marital Status with an accuracy of 0.6966 (≈ 0.7).

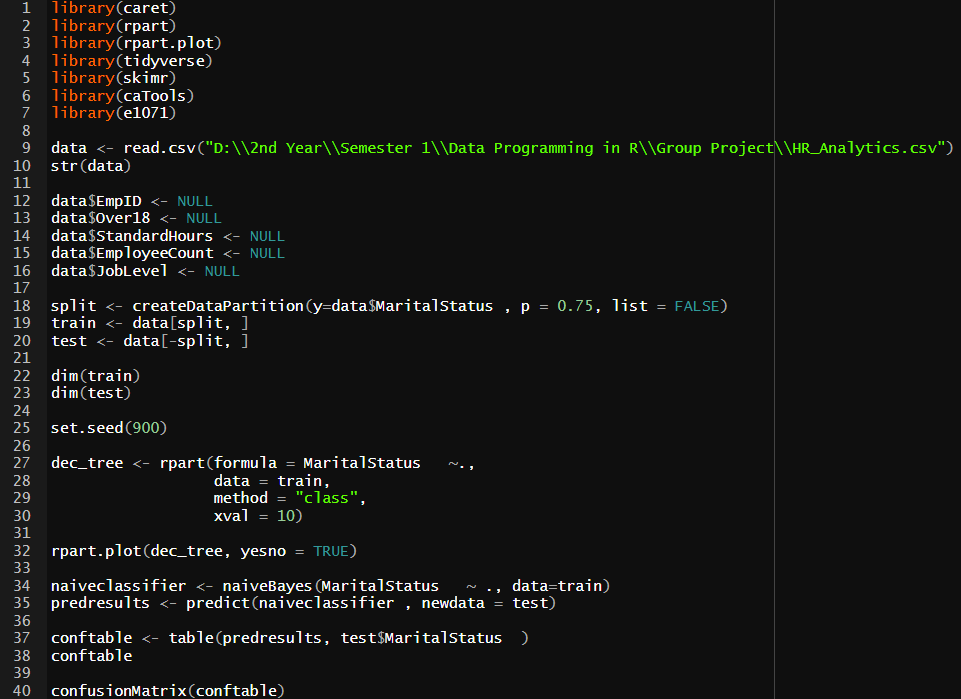
**Decision Tree Classification on Marital Status,**

A machine learning approach known as a decision tree classification creates a tree-like structure by iteratively dividing the data according to its attributes. Based on a particular feature, a decision is taken at each node of the tree, resulting in a branch that eventually generates a classification or prediction for a given input. So, we are going to generate a decision tree for the variable – Marital Status and see what kind of information we can gather.

A diagram of a relationship

Description automatically generated

The following code was used to obtain the above decision tree and create a naïve bayes model.



This was the output we got from the naïve bayes model we trained under it.

A screenshot of a computer screen

Description automatically generated

**Findings and Discussion**

\*They are written under the relevant plots and topics.