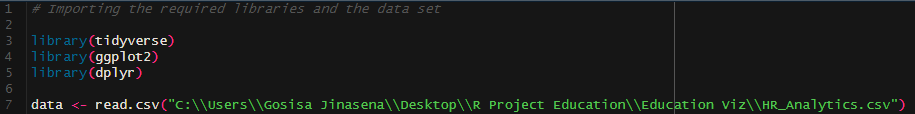
1. Education

Education is one of the main components of a HR dataset. If we take look at the dataset we can see several columns are related to education. In this part, our goal is to identify those relations and how education is affecting in other columns.

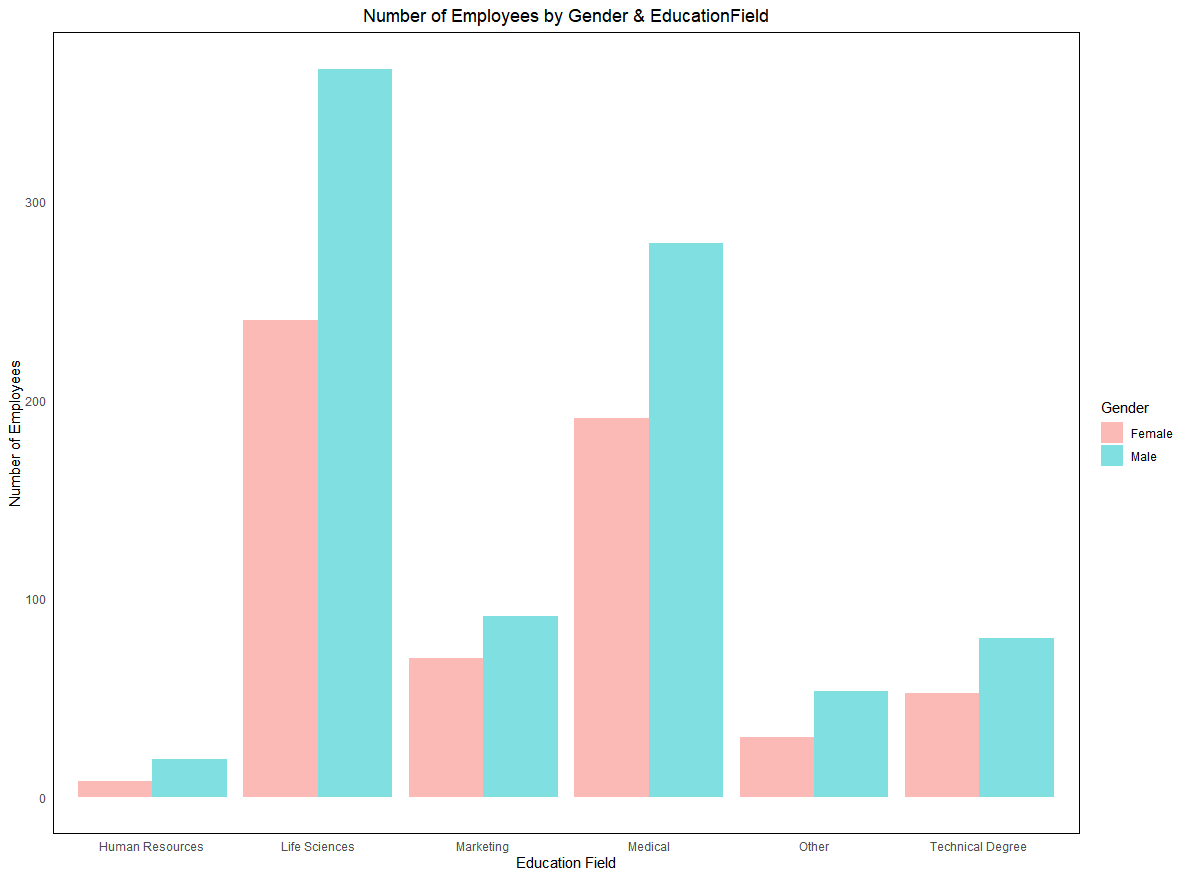
**Data Visualization**

We have got Education Field and Education Level as the columns regarding to Education. Both of them are qualitative data. So we used Bar Plots, Pie Charts, Bee Swarm Plots, Strip Charts & Violin Plots as our techniques, which are commonly used to plot qualitative data in R.

First of all we imported the required libraries and the data set to start the visualization process. Below code shows that process.

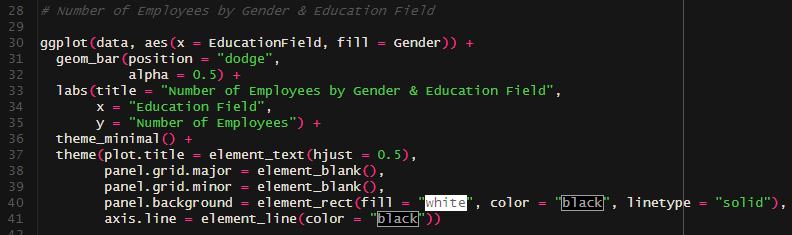


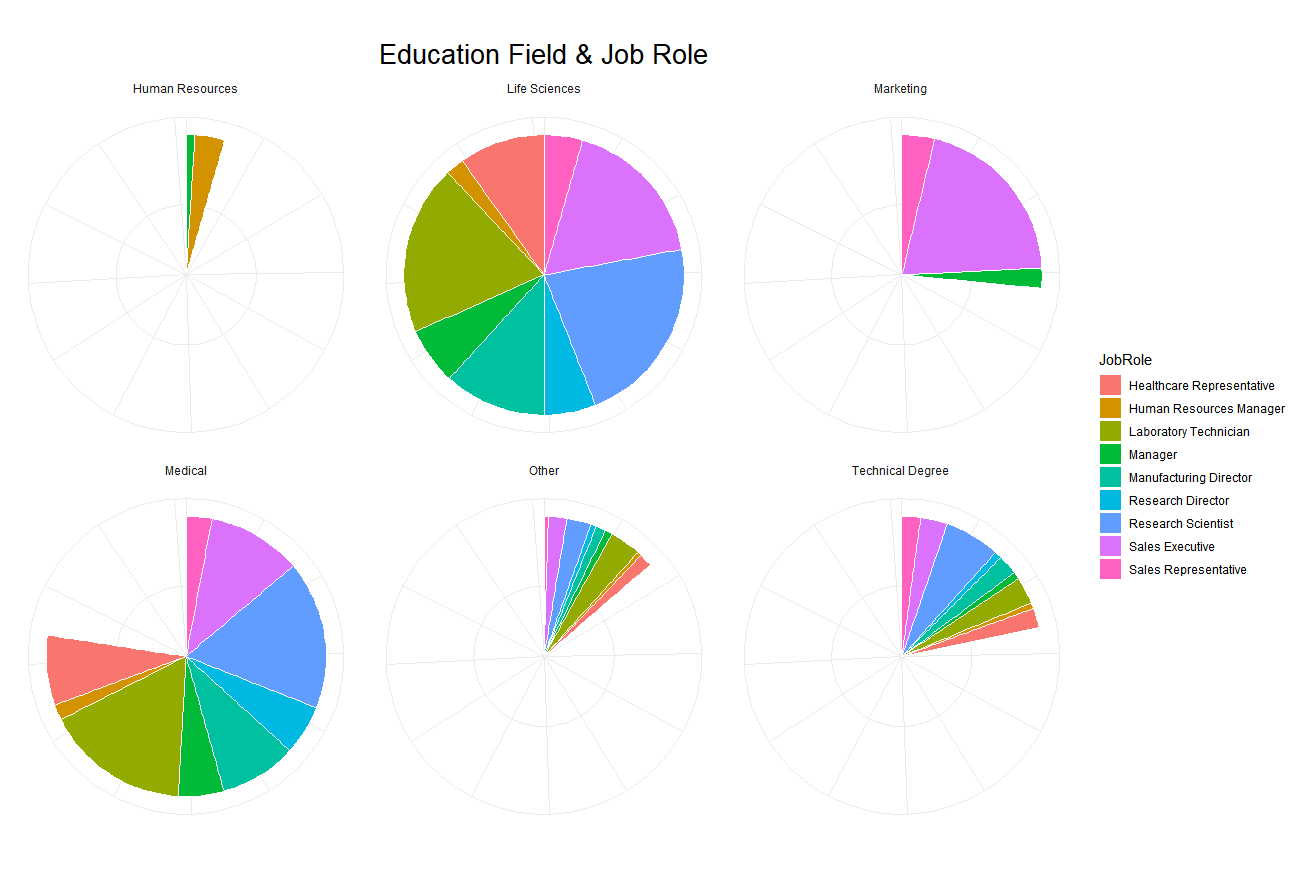
After that process we started to plot different columns with different plotting tools and methods. We are going to mention all of them below.



This bar-plot was used to check whether how employees distributed on their Education Field & Gender. According to the bar-plot Life Sciences takes a huge number of employees compared to other fields.

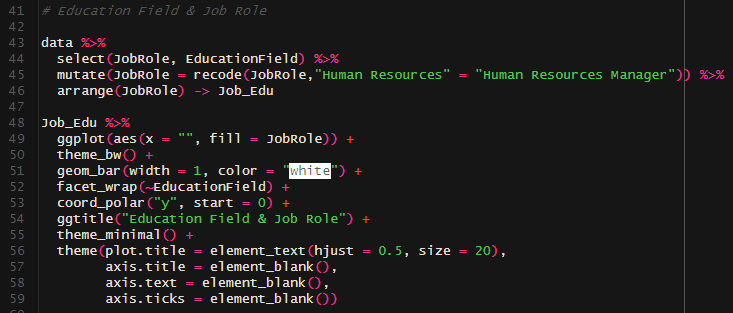
This bar-plot was generated by below code



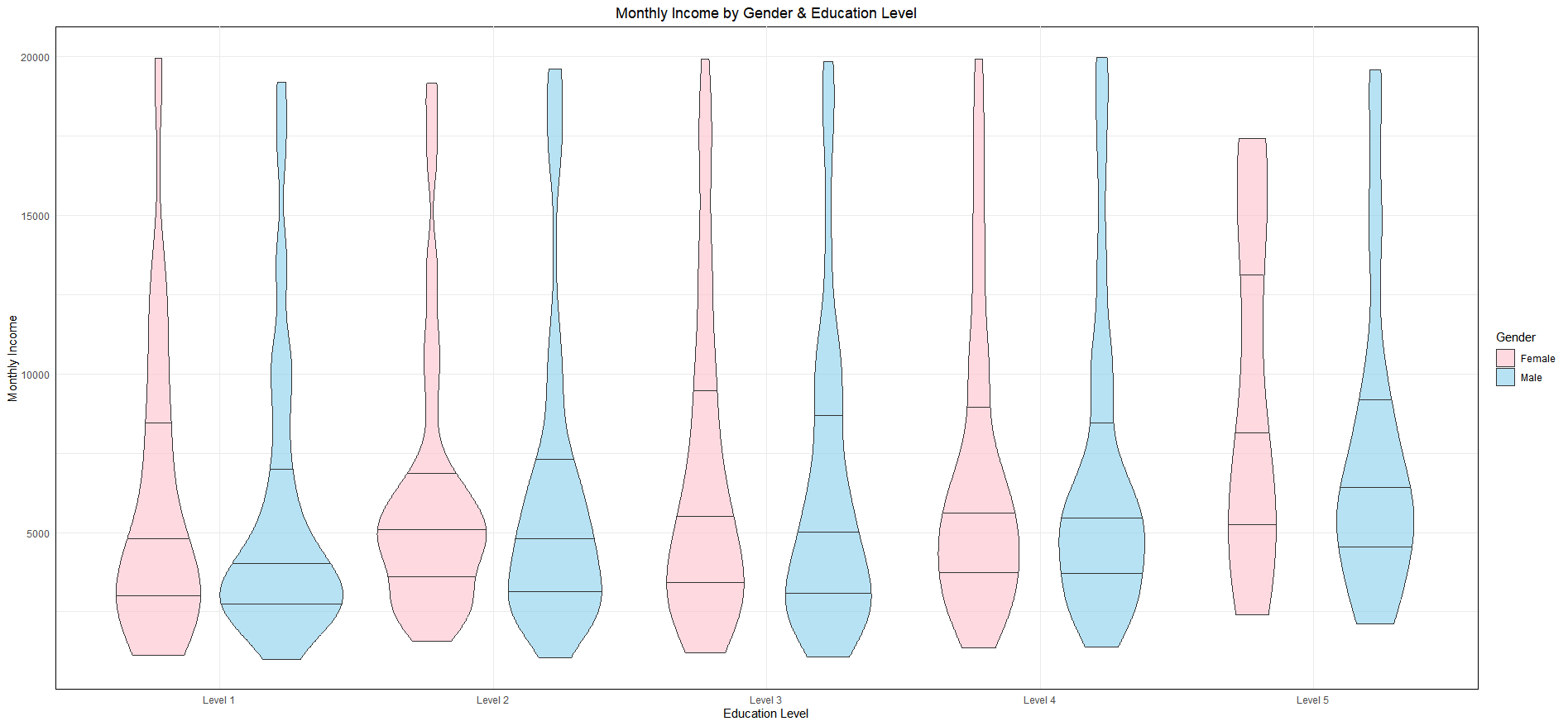


This pie charts were used to take a proper idea about how job roles are distributed with different education fields. Number of employees are shown as percentages.

Code for the pie chart is give n below

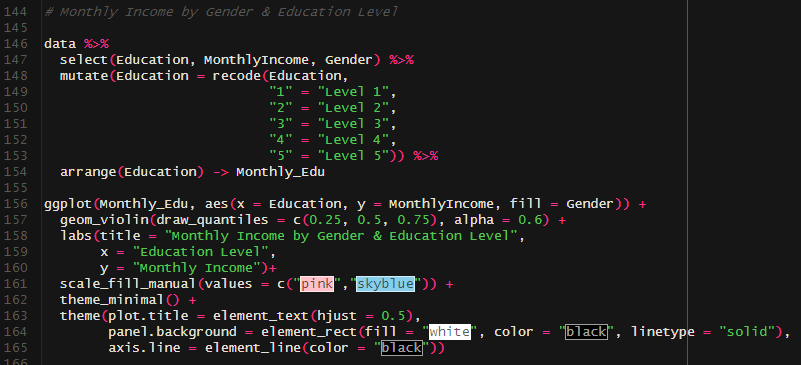


We found a job role called “Human Resources” in our dataset. Human Resources s no a job role. So, we modified that data in to “Human Resource Manager”. Top of the code snippet is mentioning the process.

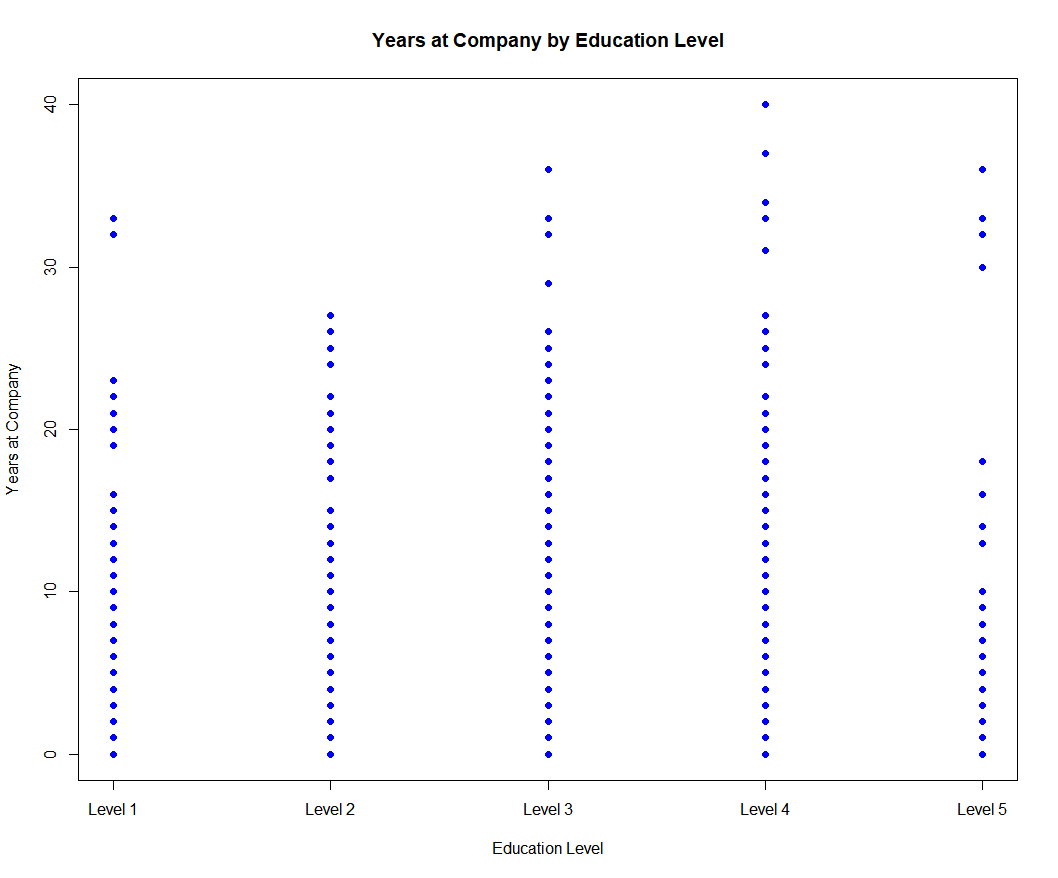


This is the violin plot for Monthly Income by Gender & Education Level. According to the 2nd quantile of the plot, we can say that levels 1,2,3, and 4 have the same behavior and each of them (50% of each level) has got around 5000 as their monthly income. Compared to the other levels level 5 has a different behavior. It has more density on the top (that means most of them have higher monthly incomes) and the beginning of monthly income for level 5 is around 2500(while other levels are beginning from 1250). According to this plot, we can consider that when the Education Level is higher level monthly income also increases relatively.

We generated this violin plot from below code.

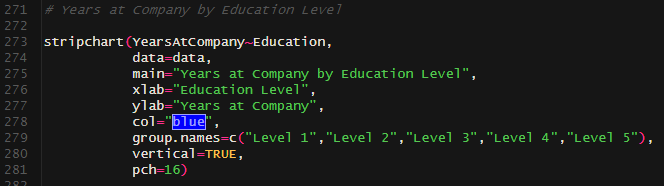


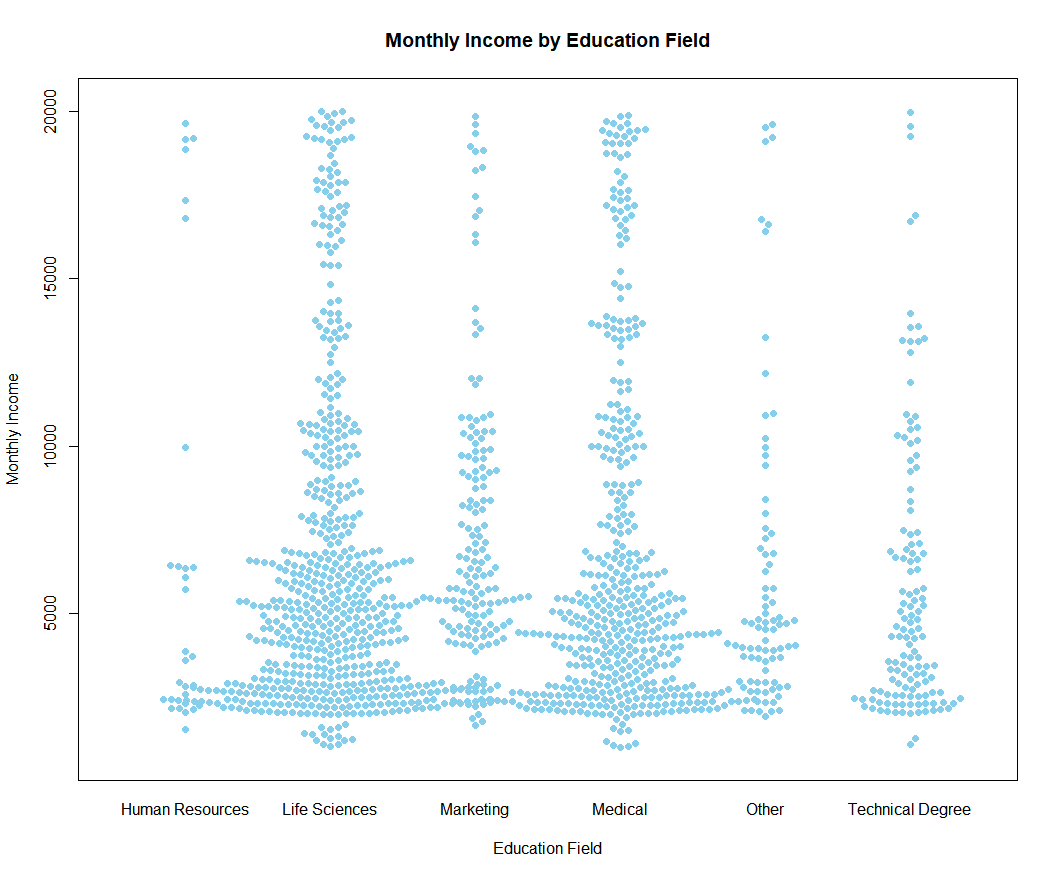
In our Education Level column, we received the levels as 1,2,3,4, and 5. For better understanding and visualization we changed them to Level 1, Level 2, Level 3, Level 4, and Level 5. At the top of the code snippet we can see the code operations.



This is the strip chart for Years at company by Education level. We can notice that the majority of employees who have level 3 and 4 level education are have worked many years at the company. It means for that employees who are at a higher education level there will be a high demand in the company.

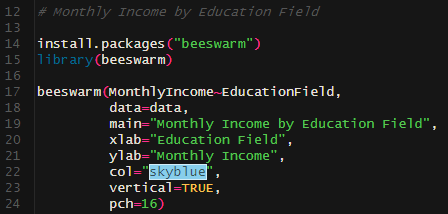
Code for the strip chart is given below



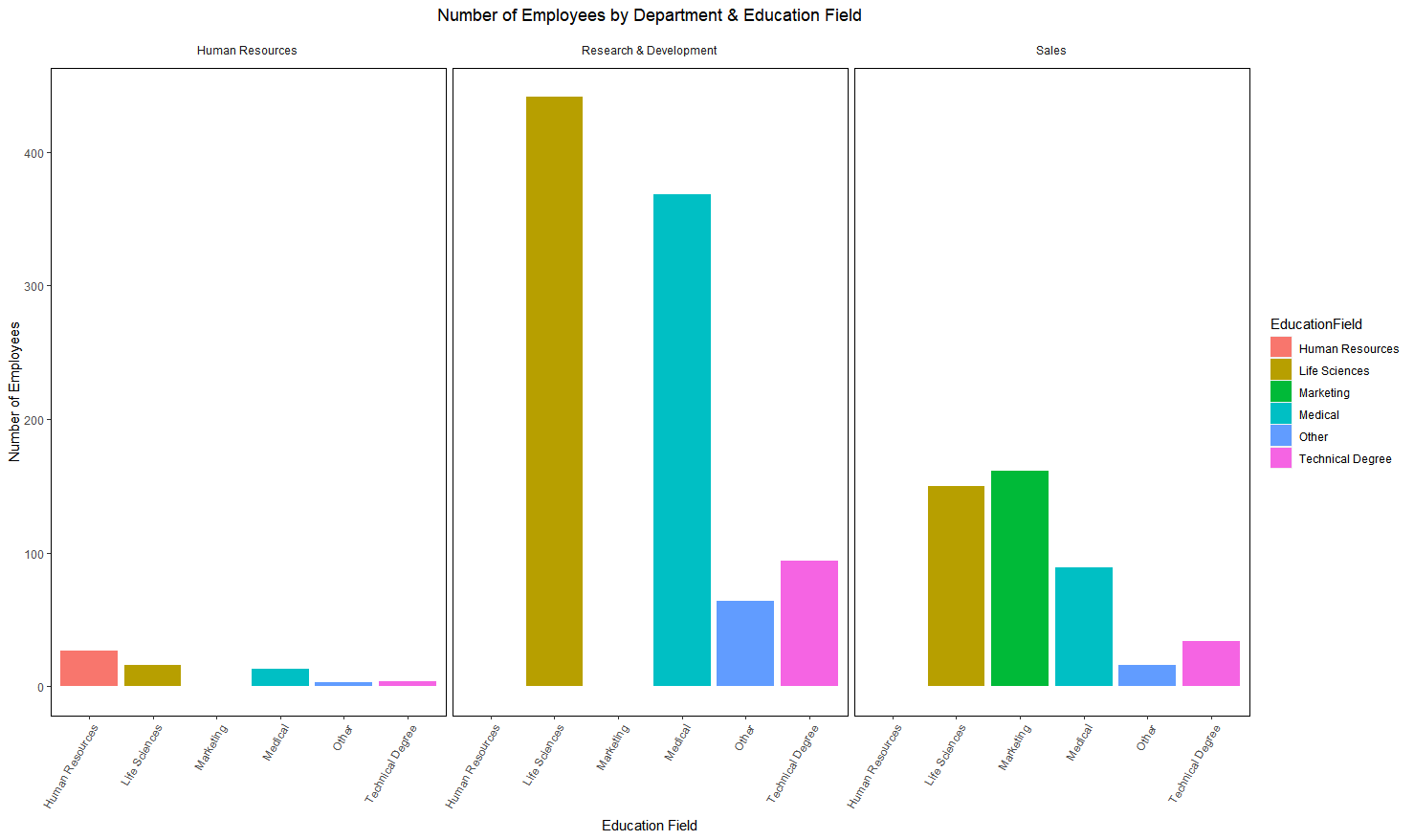


This is the bee swarm plot of Monthly income by education field. We can say that majority of employees who choose their education field as life sciences and medical are having 2500 to 7500 as their monthly income.

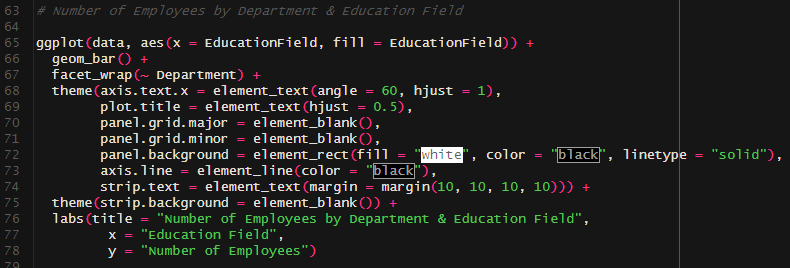
Code snippet for the above bee swarm plot is given below



We need to install the library called “beeswarm” to generate this plot.



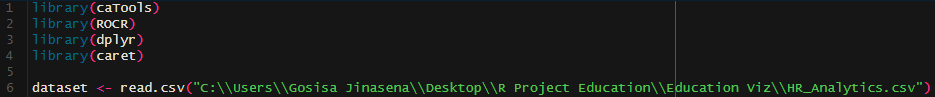
This facet wrapped bar plots were used to get an idea about how the employees distributed to the departments by their education field.

Code of the above facet wrap ped bar plots is given below

**Data Analytics through models**

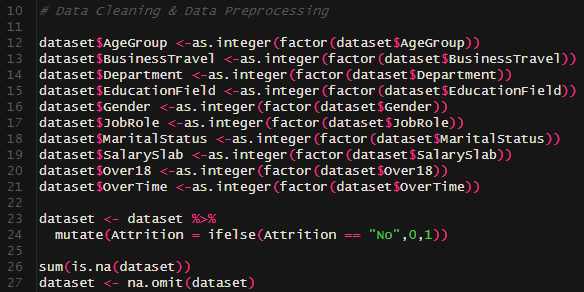
**A classification model using Logistic Regression algorithm to predict Attrition based on Monthly Income and Education Level.**

Logistic Regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome. In the data set there were two columns (Attrition & Over Time) compatible for a binary classification. So, we decided to choose attrition as our dependent variable.

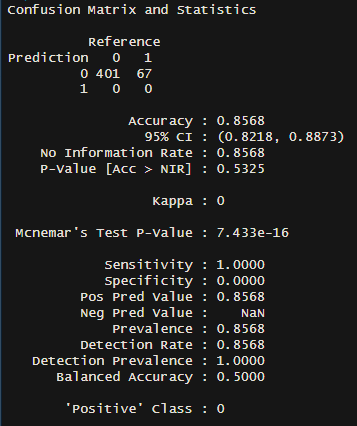


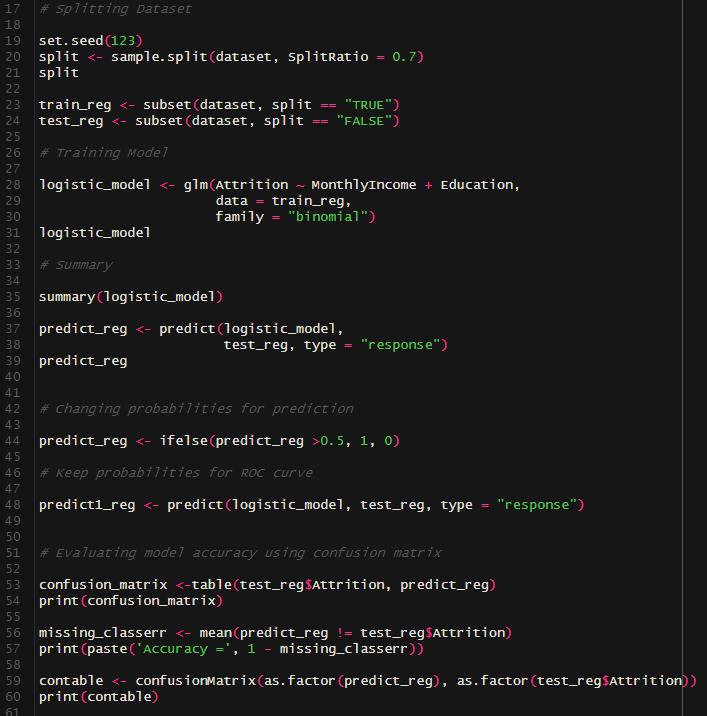
We loaded some needed libraries and the dataset to start the process. After that we started the data cleaning and the preprocessing. We checked for null values in the dataset and removed them. In logistic regression the dependent variable always must be a binary outcome. So, we converted “Yes” & “No” to “1” & “0” in Attrition column. Also converted other categorical variables into integer variables.

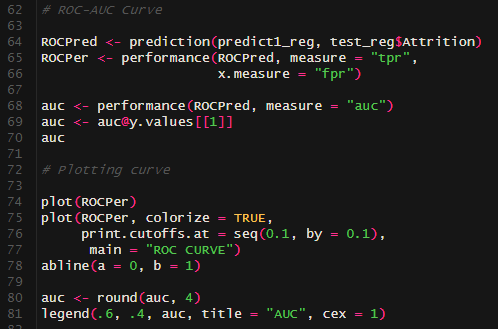
Code of the Cleaning & Preprocessing is given below

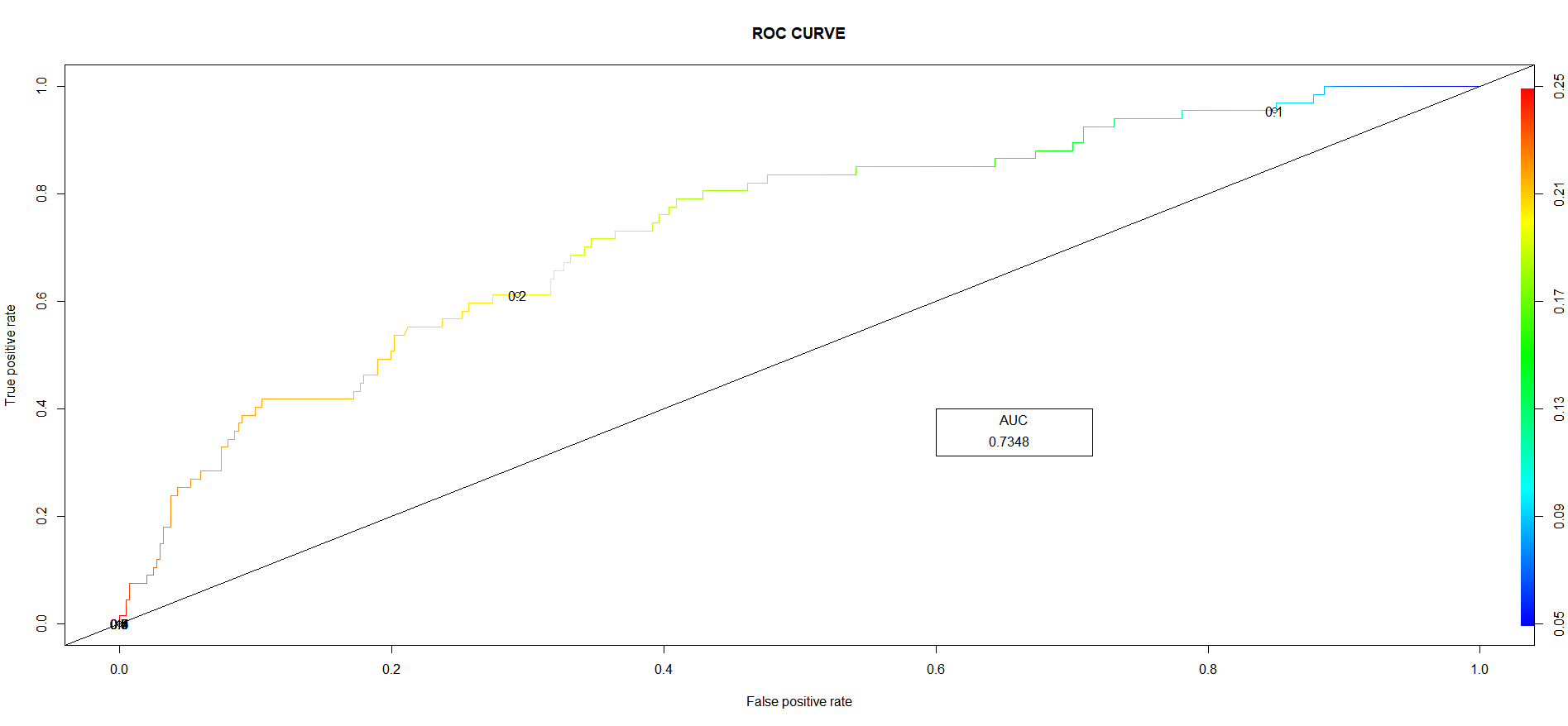


After the cleaning process done we continued to the model building process. First of all, we split the dataset in to training and testing. After that trained a logistic regression model (used attrition as dependent variable and monthly income and education level as independent variables). After that moved to the evaluation process. We used a confusion matrix for that process. It showed the model’s accuracy as 0.8568. Overall coding and the confusion matrix are given below









ROC AUC score is using to check the efficiency of a logistic regression model. We have got our AUC score as 0.7348. It is closer to 1 which means model is a good classifier.