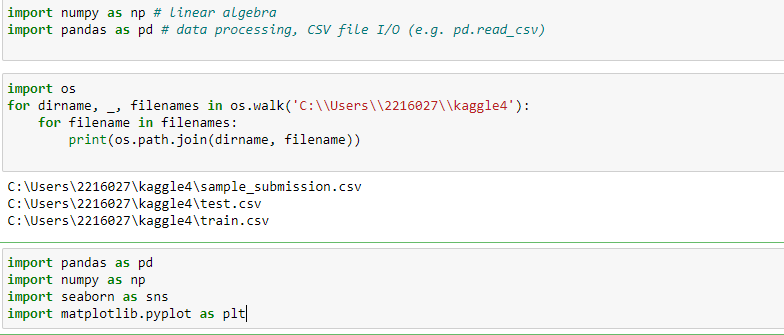
**PG\_Series\_S3\_Employement\_Retirement\_Random\_Forest**

# Created by Kaushik Kar

# Employment id- 2216027

# Importing Libraries:

First, importing the important external Python packages using the pip package manager.

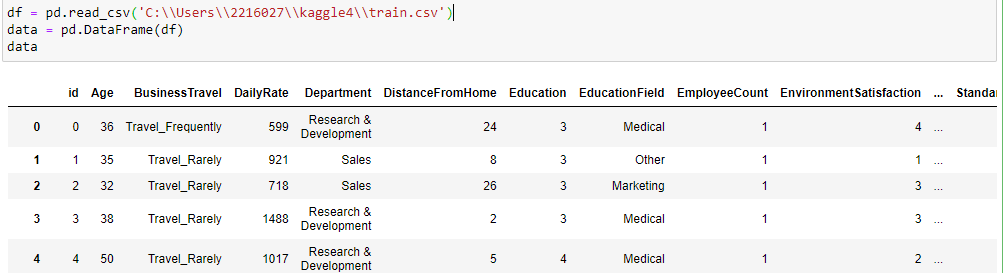


1. NumPy is used for mathematical operations like addition, subtraction, multiplication, division, etc. on arrays and matrices.
2. Pandas provides data structures for efficiently storing and manipulating large datasets, and tools for reading and writing data to and from various file formats, including CSV, Excel, and SQL databases
3. Seaborn is a data visualization library based on Matplotlib which is a plotting library used for creating static, interactive, and animated visualizations in Python.

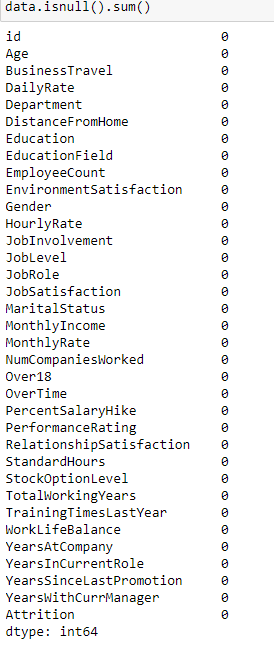
This above code is using the OS module to walk through a directory tree and print out the path of each file in the tree such as test.csv, train.csv and sample\_submission.csv

1. **Data Import:**

With the help of Pandas library, we can read and upload the data in csv form.



* **Null Values:** isnull() is the process by which we can check whether there is any null value present in the data or not i.e. isnull() is part of the data cleaning stage. Here in the data, we found that there is no missing values as shown in the below image.



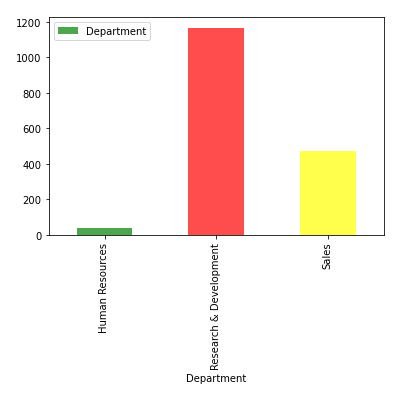
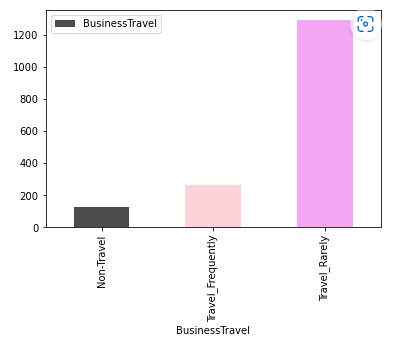
info and describe are part of the exploratory data analysis stage. info is used to display a concise summary of a data frame, including the number of non-null values and the data types of each column. This method is often used in the exploratory data analysis stage to gain an initial understanding of the structure and contents of a data frame.

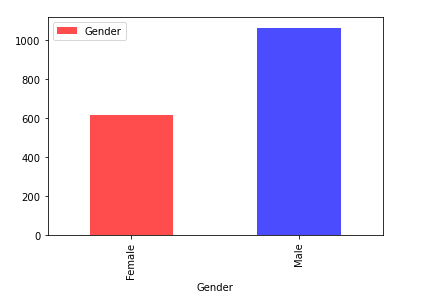
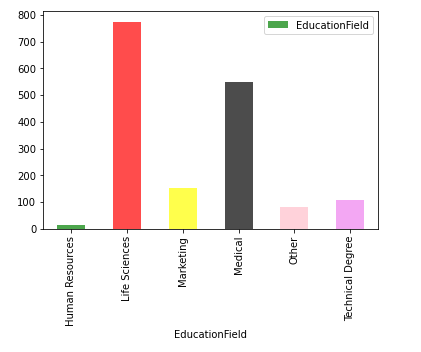
describe is used to generate descriptive statistics of a data frame, such as the mean, standard deviation, and quartile ranges of numeric columns. This method is often used in the exploratory data analysis stage to understand the distribution and range of values in the data.

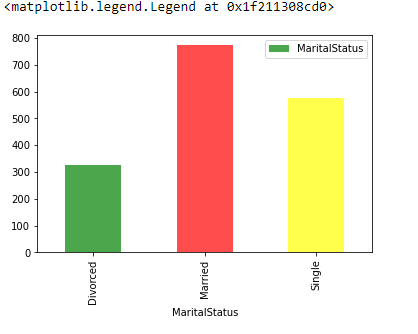
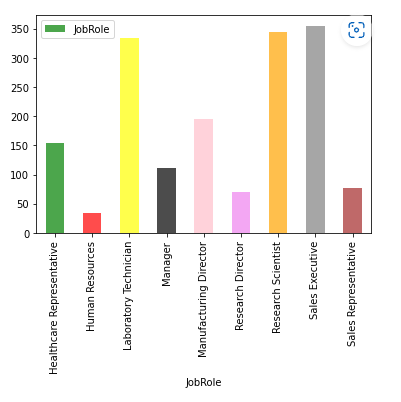
1. **Grouping and Visualization:**

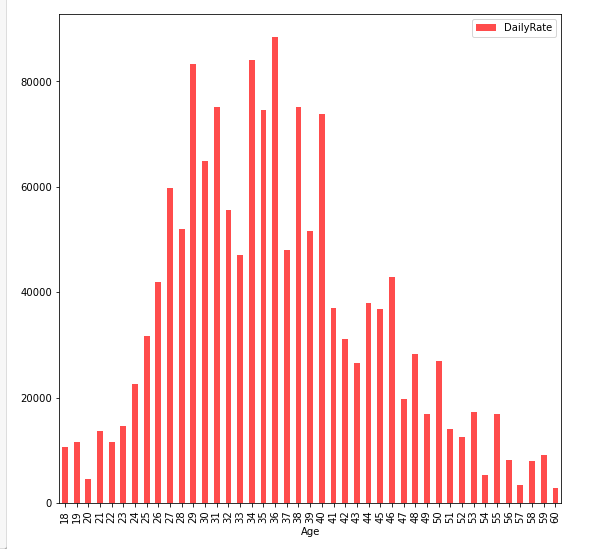
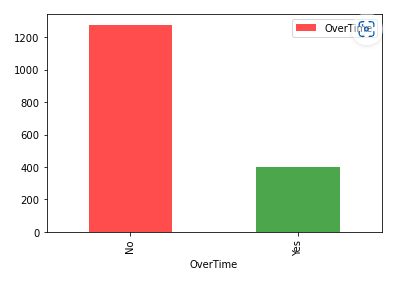
In the data there are categorical columns called as Business Travel, Department, Education Field, Gender, job role, marital status, overtime, Age with Daily Rate.

We can group by all these categorical columns by itself and make plot of each column so that we can find the categories for each column through visualization.









Now uploading the test.csv data and try to predict the model through Random Forest.

**Building a forecast model and getting the result:**



* The train\_test\_split function is used to split a dataset into training and testing subsets for machine learning. This is a common step in the machine learning workflow, as it allows us to evaluate the performance of our model on data that it hasn't seen before.
* Now we drop the column “Attrition” and consider this with the variable x, these are the features and Consider Attrition as a y variable which is the target variable.
* The RandomForestClassifier is an ensemble learning method for classification that fits multiple decision trees on different sub-samples of a dataset and aggregates the results to make a final prediction.
* The RandomForestClassifier instance r is created with hyperparameters n\_estimators=30, oob\_score=True, max\_features=5, and max\_depth=30. The model is then fit to the training data using r.fit(x, y) and the predicted class labels of the testing data are stored in y\_pred using r.predict(test).
* RandomForestClassifier model predicted the target variable with an accuracy of approximately 99.94% on the training data.