Summer Internship Program

Henry Harvin Education India LLP

Name: Swarnabh Srivastava

Course: Summer Internship Program (SIP) Python

Batch: Jun-Jul 2019

Job: Business Analyst Associate (Intern)

Institution: Narsee Monjee Institute of Management Studies

# 

# DECLARATION

I hereby declare that the project report entitled “**HR Analytics**” submitted by me to **HENRY HARVIN EDUCATION INDIA** is a record of bonafide project work carried out by me. This project is an original report with references taken from websites and help from mentors and teachers.

DATE: 28 Jul 2019

Swarnabh Srivastava

SIP – Python

# TABLE OF CONTENTS

Declaration 1

1. Project Introduction 3

2. Project Data Introduction 7

3. Exploratory Data Analysis (EDA) 8

4. Model Building 18

5. Prediction and Evaluation 22

Bibliography 29

deeper insight into behavioral particularities and characteristics of its employees.

# 

**Overview** :

In the given project we will analyze the data given by checking the various variables and their dependency with the prediction and calculate them.

There are many variables which are depending on each there and we have to find the dependency within each other.

# Project Data Introduction

This project is based on Predictive Analysis. This is a Python-based Project. This project was created via Spyder 3.3.5. IDE (Integrated Development Environment) using Python 3.7.3 and Ipython Console 7.4.0. The final outcome of this project is saved as a Jupyter Notebook v7.8.0. The libraries of python used in this project are:

1. NumPy

2. Pandas

3. Matplotlib

4. Seaborn

5. Statsmodels

6. Sci-kit Learn

This project is based on a data set provided by the teachers via GITHUB. The data used in the project is discrete and categorical, and hence, we are using LOGISTIC REGRESSION and RANDOM FOREST CLASSIFICATION for predicting our data.

Here, the **target variable** is LEFT.

Data Set Dictionary:

|  |  |  |
| --- | --- | --- |
| **Name of Column** | **Description** | **Type** |
| satisfaction\_level | Satisfaction level of the employee | Numeric |
| last\_evaluation | Last evaluation of the employee | Numeric |
| Number\_projects | No of projects completed by the employee | Categorical |
| Average monthly hours | Average Monthly hours spent by the employee | Categorical |
| Time spent in company | Time spent by employee in the company | Categorical |
| **Name of Column** | **Description** | **Type** |
| Work accident | Accident happened to employee or not while working | Categorical |
| Left (target variables) | Employee left or not | Categorical |
| Promotion last 5 years | Promotion status in the last 5 years | Categorical |
| Department | Department of the employee | Categorical |
| Salary | Salary category | Categorical |

Data Set Size: 9653 rows and 10 columns

**Categorical Variables:**

[number\_projects, work\_accident, left, promotion\_last5years, department, salary, time\_spend\_company, average\_monthly\_hours] = 8 features

**Numeric Variables:**

[ satisfaction\_level, last\_evaluation] = 2 Features

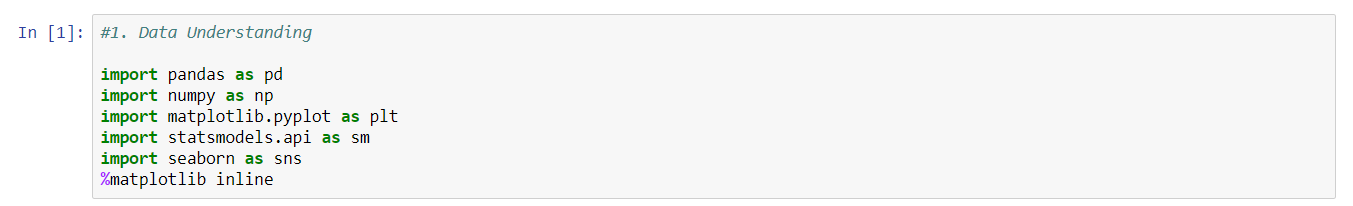
# 3. Exploratory Data Analysis (EDA)

In statistics, **exploratory data analysis** (**EDA**) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by many to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

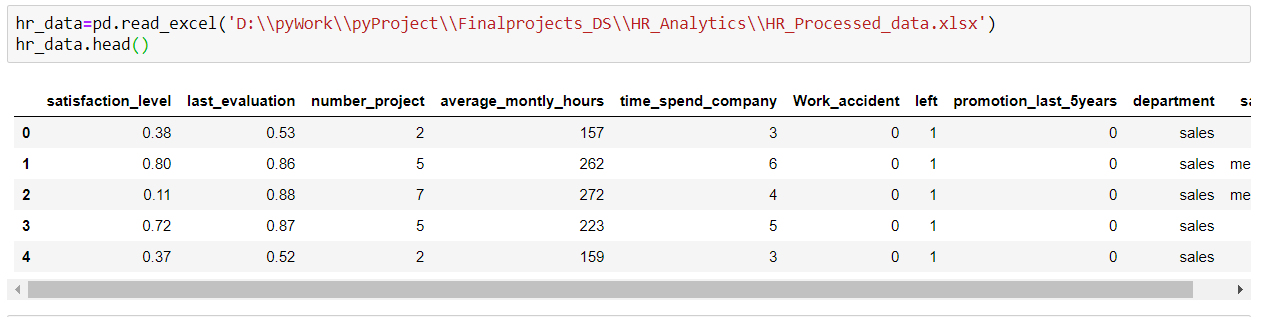
In this project, we used matplotlib, seaborn for EDA using python 3.7.3. It is as follows:

## 3.1. Data Understanding

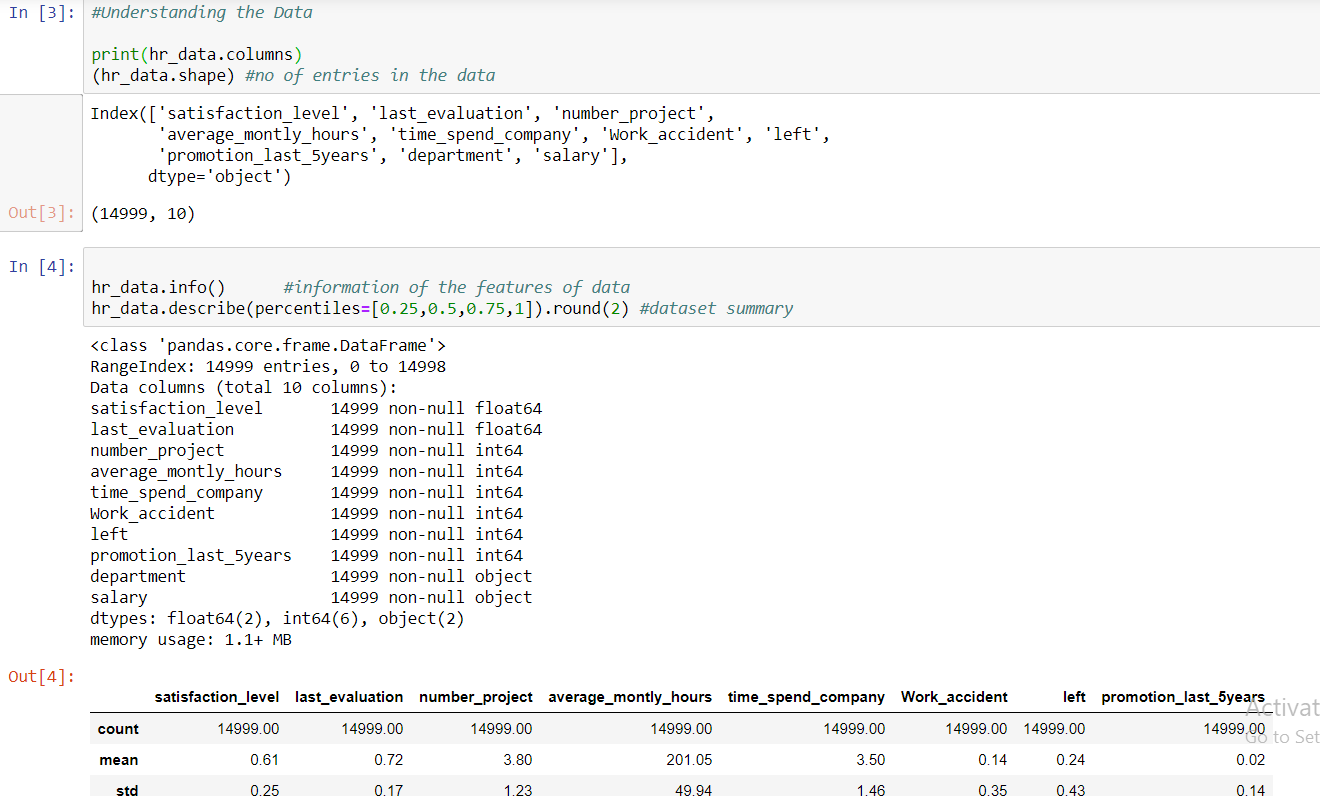
Importing all the necessary libraries



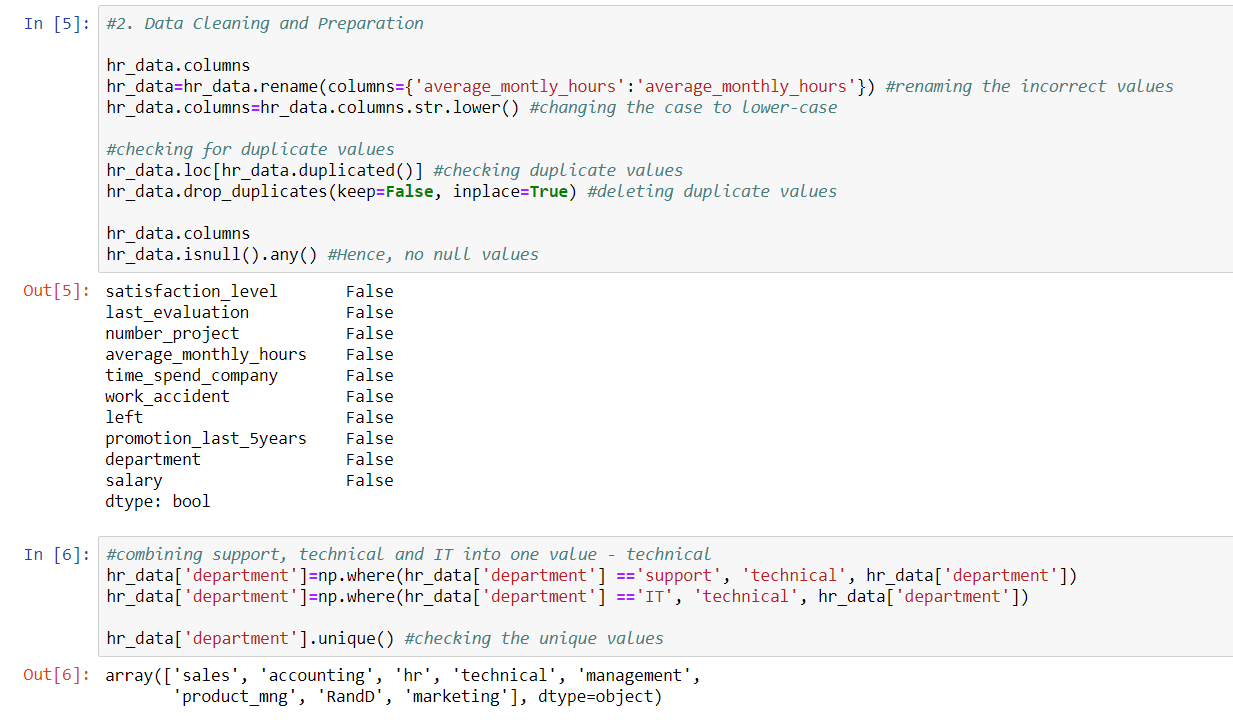
Importing the dataset



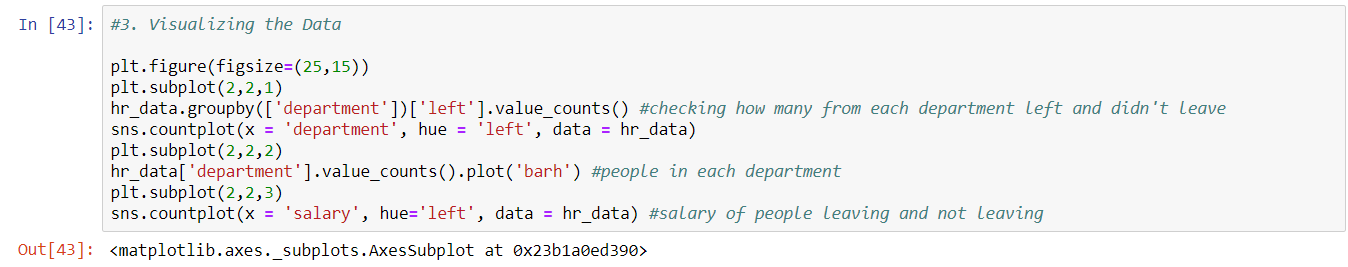
The **describe()** function and **info()** function to study the data set.

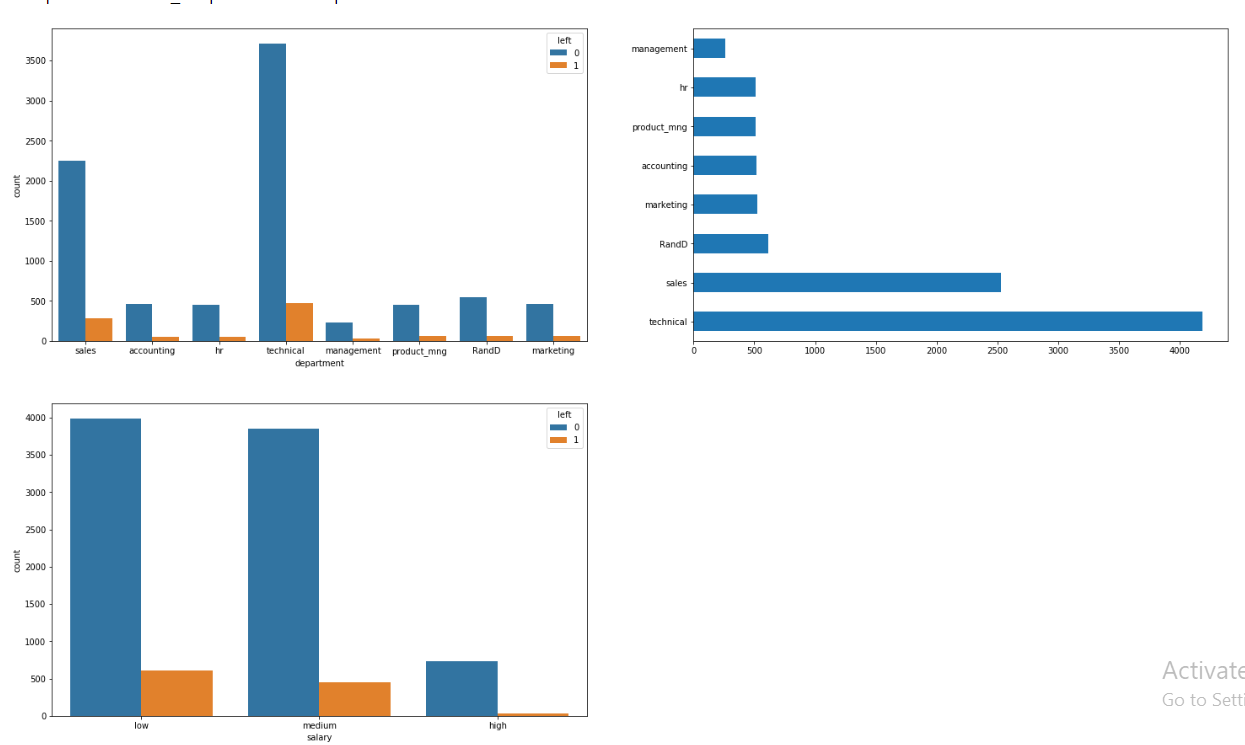


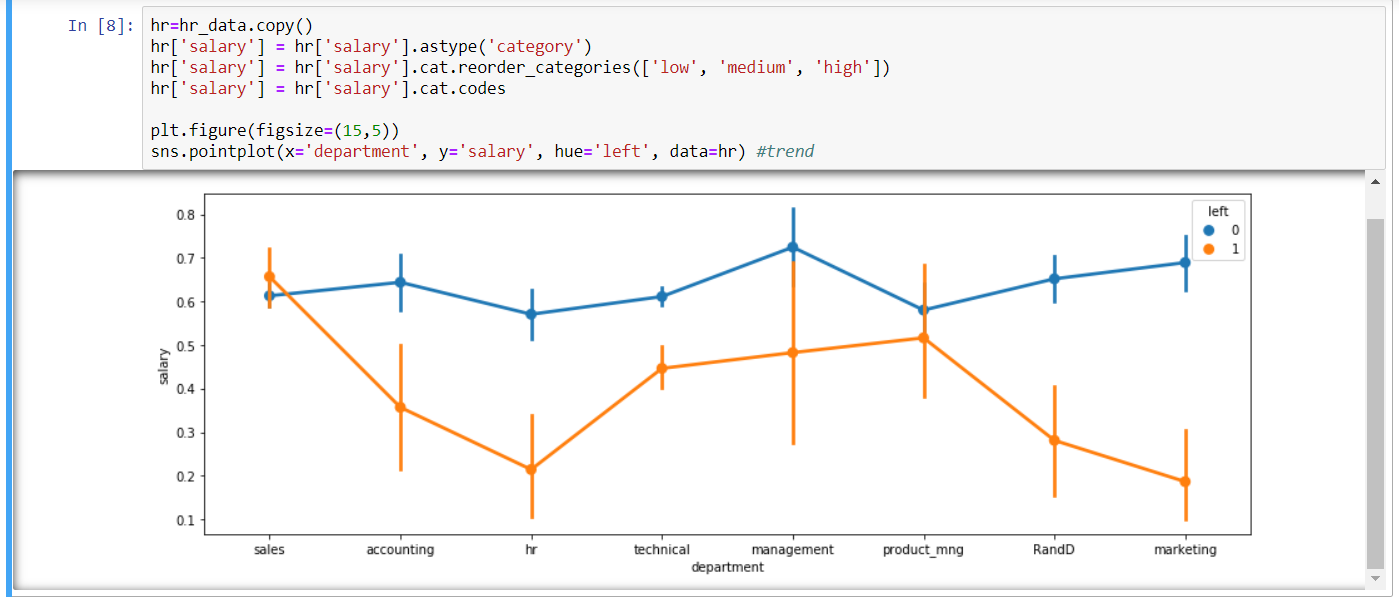
## 3.2. Data Cleaning and Preparation

Our data is now clean. It’s time for the final step of EDA: Visualization.

## 3.3. Visualization

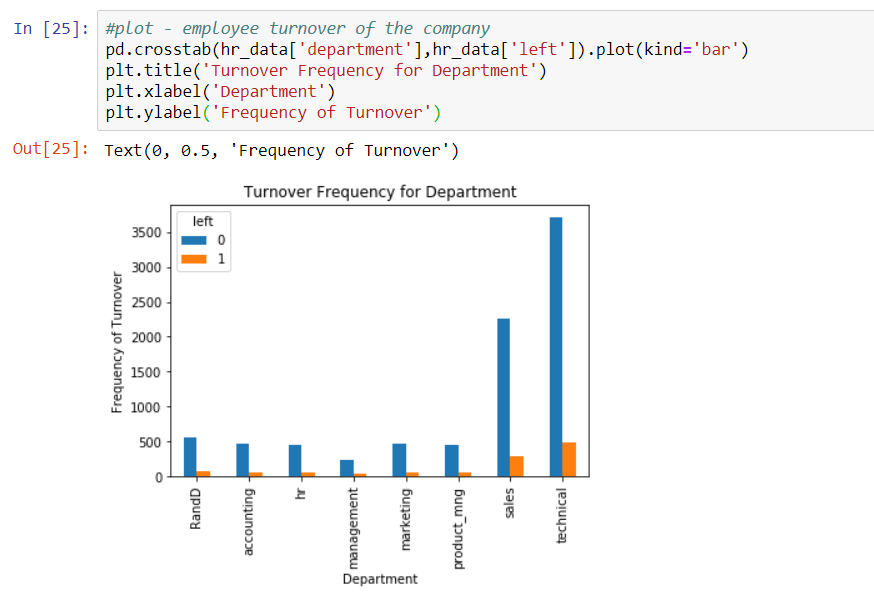




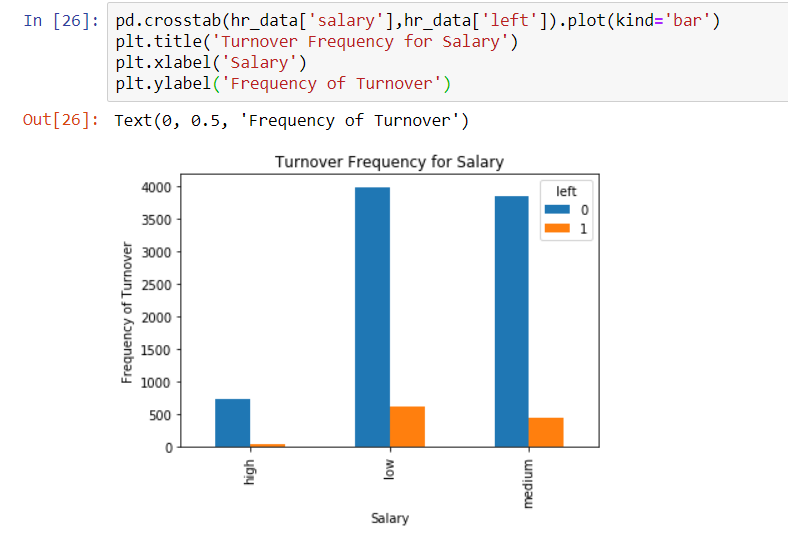


**Categorical Variables:** [number\_projects, work\_accident, left, promotion\_last5years, department, salary, time\_spend\_company, average\_monthly\_hours] = 8 features

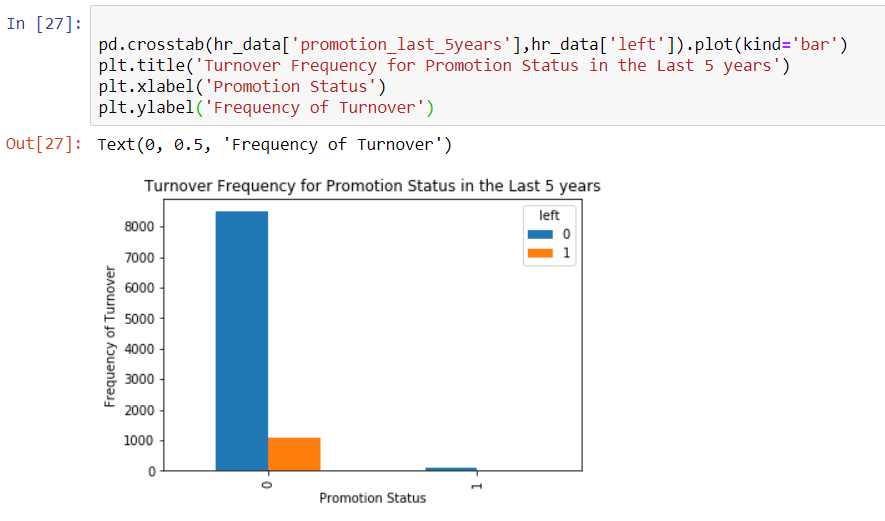
**1. Department**

This shows that people are leaving the most from *technical.* Hence, department is an essential variable.

**2. Salary**

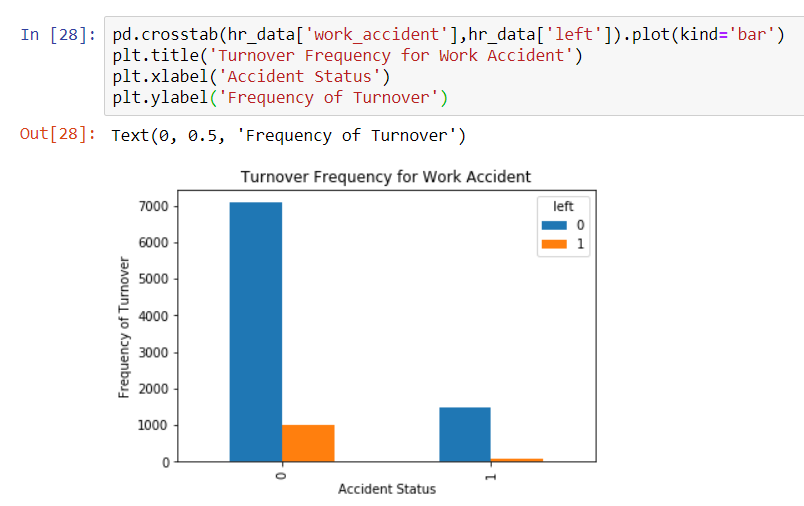
This clearly shows that medium salary and low salary people are leaving more frequently.

**3. Promotion Last 5 Years**



This shows that people who didn’t get a promotion in the last 5 years were the most to leave.

**4. Work Accident**

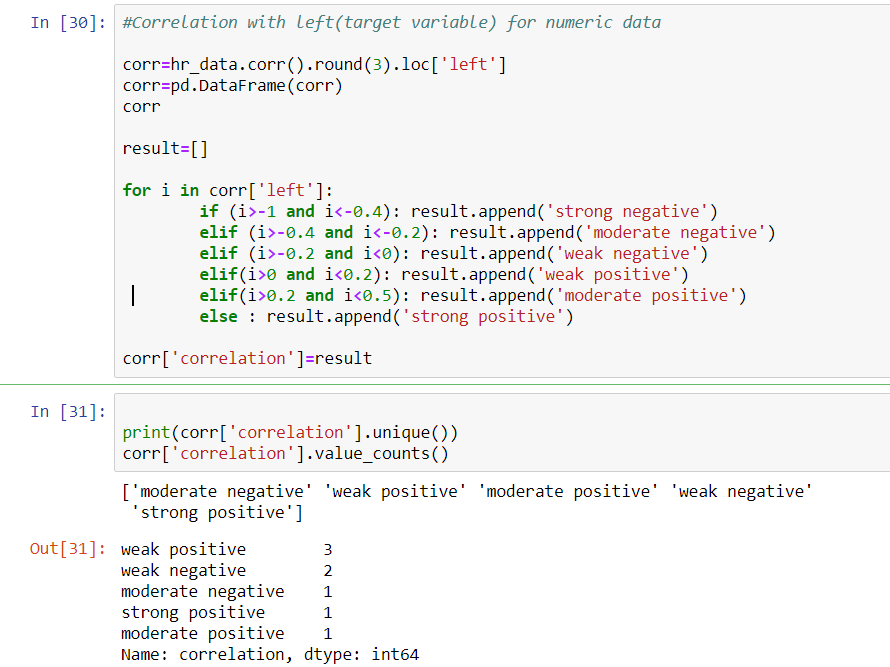
From the graph, we can say that the employees that were not involved in a work accident were leaving more frequently.

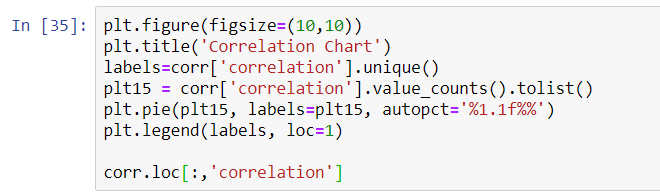
**5. Number of Projects , Time Spent in Company**

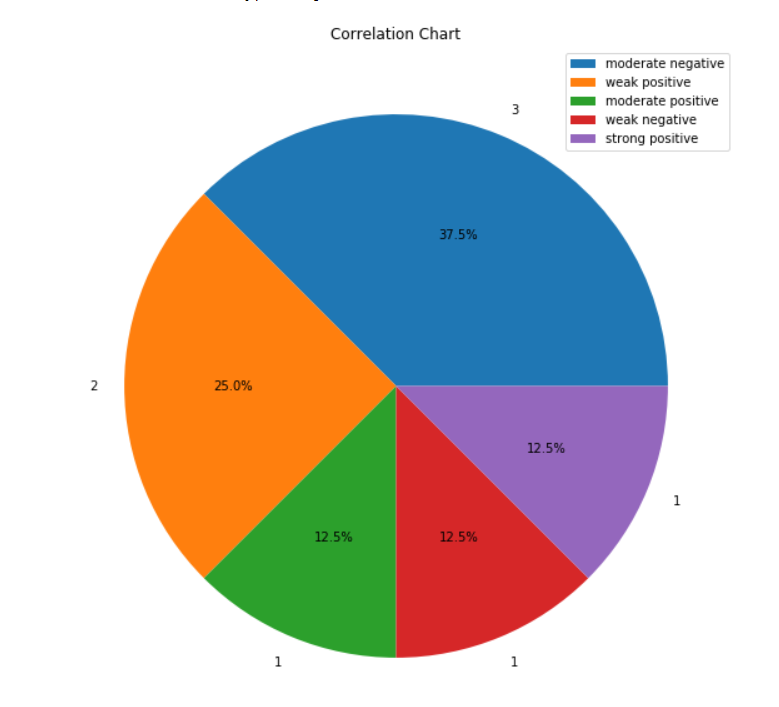


This shows that number of projects 4 were completed most frequently. Also, most employees spent 3 hours in the company.

With this conclusion, **the categorical variables visualization is complete**. Now, let us make a correlation data frame for *left*.







This correlation chart shows that most of the categories are moderately positively correlated with the target variable (‘left’).

After the visualization, the variables selected for analysis are:

1. Satisfaction Level

2. Time Spend Company

3. Last Evaluation

4. Number of Projects

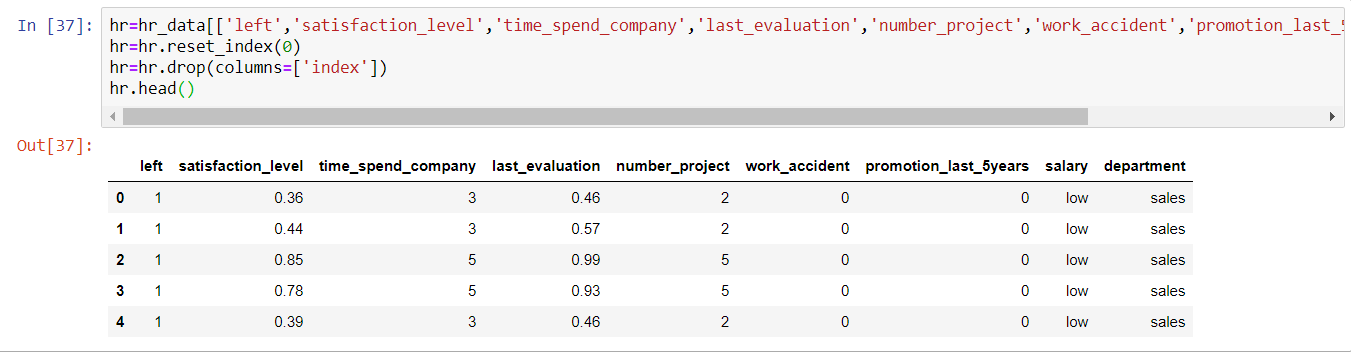
5. Work Accident

6. Promotion last 5 years

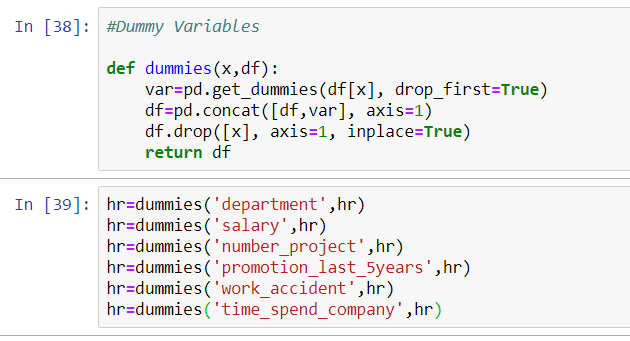
7. Salary

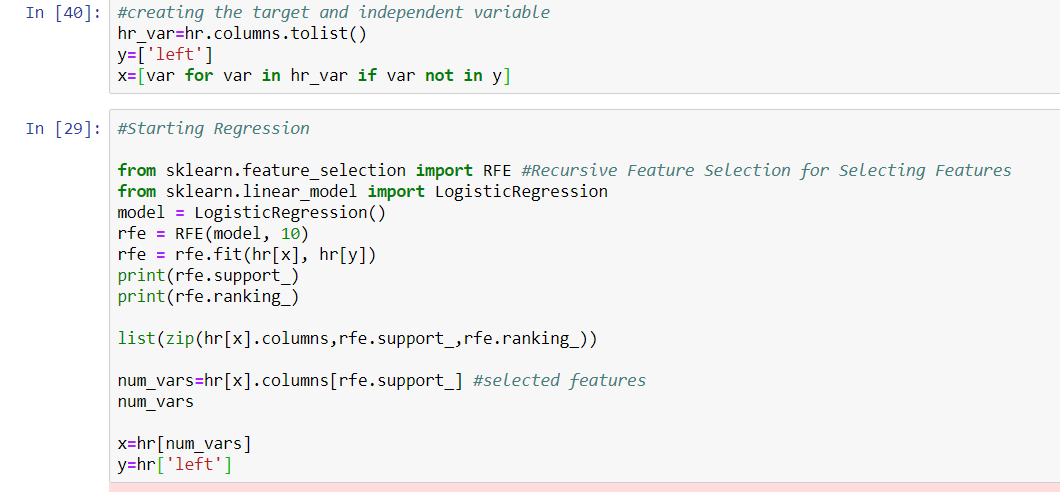
8. Department

# 4. Model Building

 **Dummy variables**

We need to create dummy variables for categorical variables only.





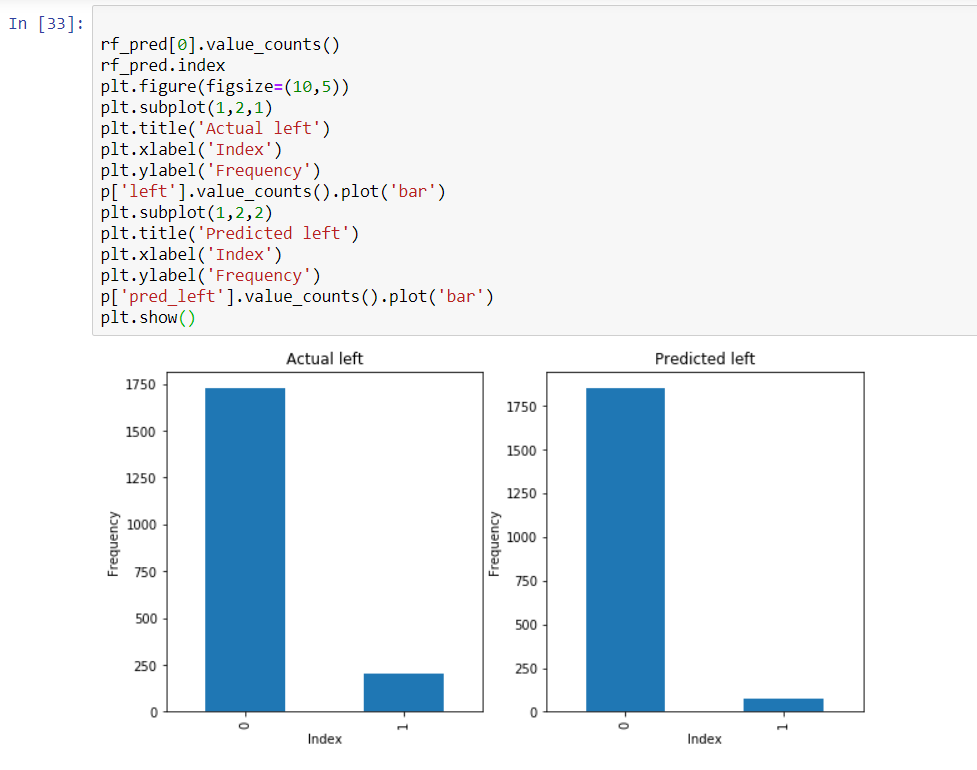
We are using logistic regression because our target variable is categorical or binary (0,1) and is not continuous.

# 5. Prediction and Evaluation

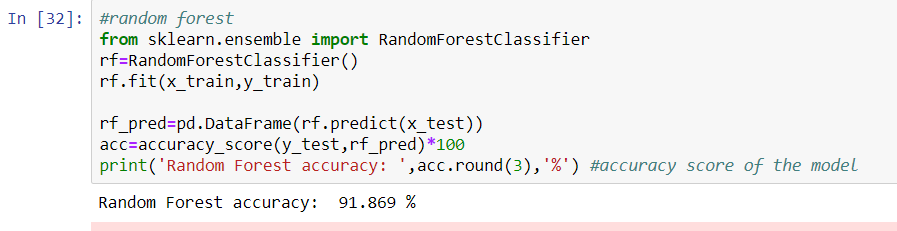
We are at the final stage of our project. Time to predict values!



**Our model accuracy is 88.607%.** It’s perfect for logistic regression. Now, let us check the actual and predicted values.



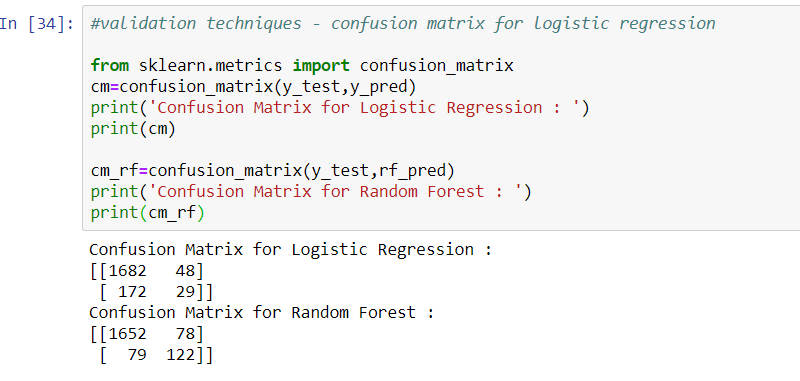
Random Forest Classifier for our analysis.



**The random forest accuracy is 91.869%.**

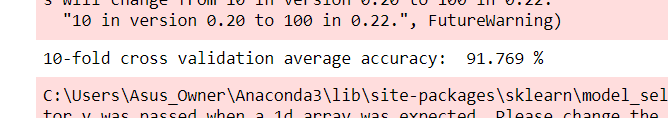
This certainly means that random forest is a better method for this data. Let us validate our analysis using *confusion* *matrix* and k-fold cross validation.

**Confusion Matrix**:

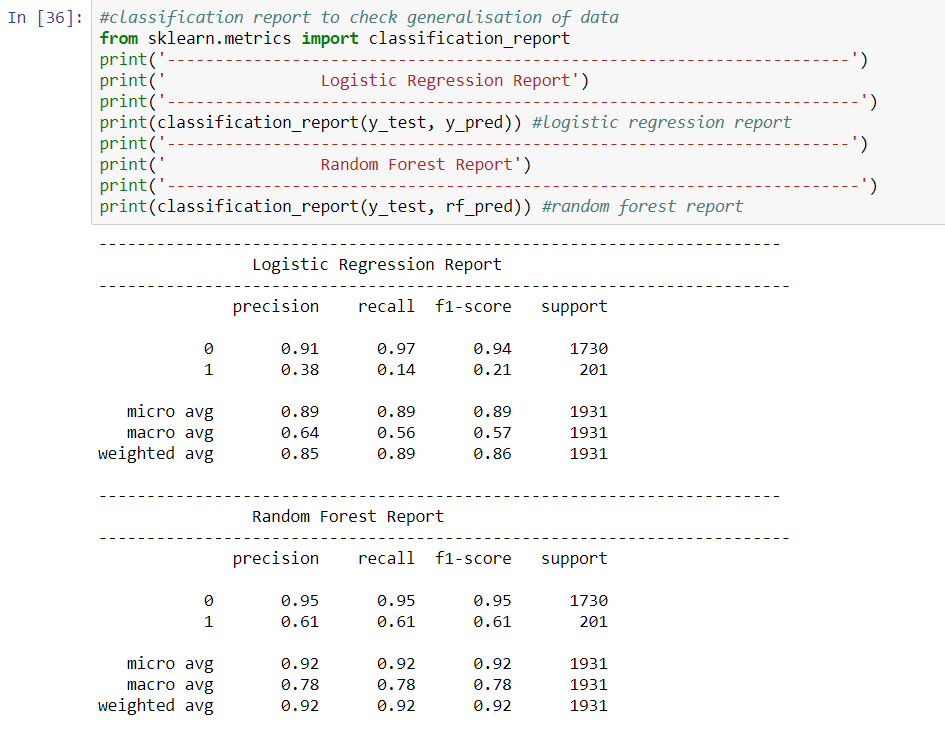


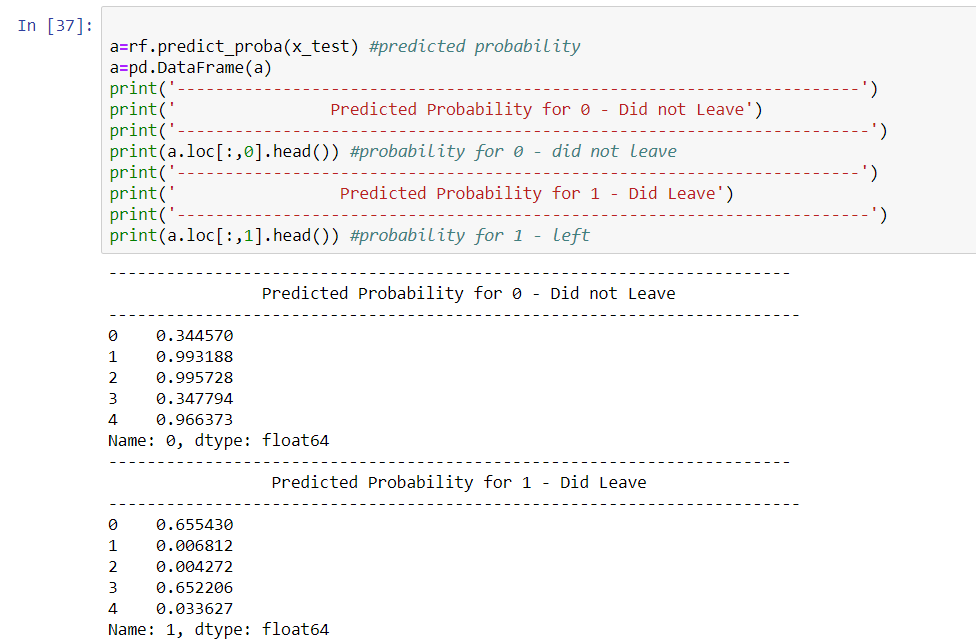
**K-Fold Cross Validation**:

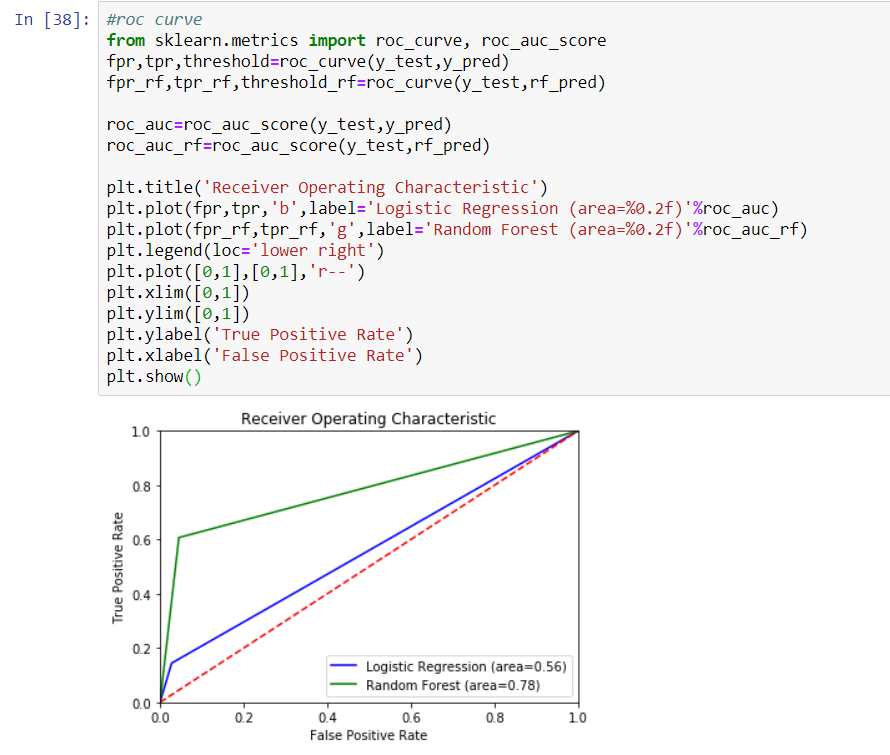




We can see that **our cross validation accuracy is 91.769%**, which is almost as much as our random forest accuracy. Hence, we can say that there is **negligible overfitting** of our model.





Here, we can infer that Random Forest is better than Logistic Regression in this case.

# *Bibliography*

* <https://www.cgi.com/sites/default/files/hr_whitepaper.pdf>
* <https://www.academia.edu/39249197/HR_ANALYTICS_A_MODERN_TOOL_IN_HR_FOR_PREDICTIVE_DECISION_MAKING>
* <https://en.wikipedia.org/wiki/Logistic_regression>
* <https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>
* <https://machinelearningmastery.com/k-fold-cross-validation/>
* <https://github.com/akjadon/Finalprojects_DS/tree/master/HR_Analytics>