

***HUMAN RESORCE DATA ANALYTICS USING PYTHON***

**Mentor-Tania Chakraborty**

***Done by:***

***Pushpendu Nath.***

***Amarjit Ghosal.***

**Acknowledgements**

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them.

We are highly indebted to Ms. Tania Chakrabortyfor their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

We would like to express my gratitude towards our parents & member of NIVT Skills for their kind co-operation and encouragement which help me in completion of this project.

We would like to express our special gratitude and thanks to industry persons for giving me such attention and time.

"You don't build a business. You build people, and people build the business."

- Zig Ziglar.



**What Is Human Resources?**

"Human resources" (HR) is a term that is used in business to refer to the people who work for a company or organization. It also is used to refer to the department of a company that is responsible for managing those resources, such as hiring and training new employees and overseeing the benefits and compensation packages provided to all of the company's employee.

**What Is Human Resources Analytics?**

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment.



**What Human Resources Analytics Does?**

As many human resources experts know, what human resources analytics actually does is demonstrate the impact that an HR department has on the organization in which it is embedded. In short, HR analytics demonstrates the causal relationship between the activities exacted by an HR department and the business outcomes that result from this activity. Once this causal relationship is discovered, HR analytics professionals will devise and implement a strategic plan to assist the business in attaining better outcomes.

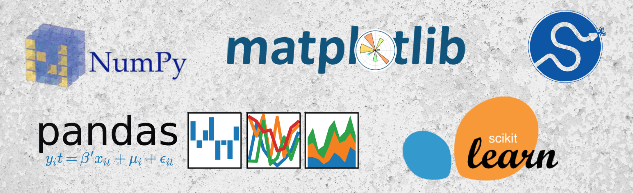


**Why We Use The Python Language To HR Analytics?**

Python's expansive library of open source data analysis tools, web frameworks, and testing instruments make its ecosystem one of the largest out of any programming community.

Python is an accessible language for new programmers because the community provides many introductory resources Programming languages have unique ecosystems, cultures and philosophies built around them. You will find friction with a community and difficulty in learning if your approach to programming varies from the philosophy of the programming language you've selected.

Python's culture values open source software, community involvement with local, national and international events and teaching to new programmers. If those values are also important to you and/or your organization then Python may be a good fit.

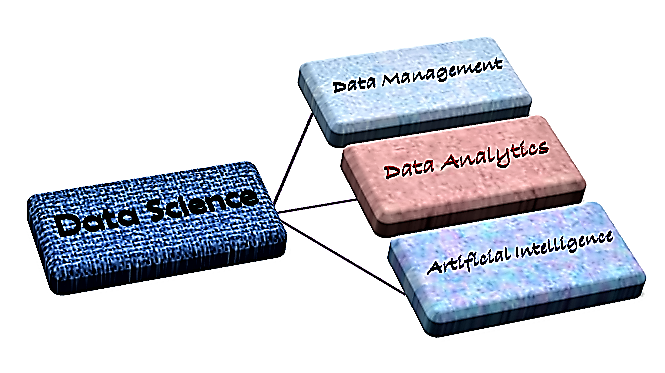


**What Is Data?**

In computing, data is information that has been translated into a form that is efficient for movement or processing. Relative to today's computers and transmission media, data is information converted into binary digital form. It is acceptable for data to be used as a singular subject or a plural subject. Raw data is a term used to describe data in its most basic digital format .

**What Does Data Science Mean?**

Data science is a multidisciplinary blend of datainference, algorithmm development, and technology in order to solve analytically complex problems. This aspect of data science is all about uncovering findings from data. Diving in at a granular level to mine and understand complex behaviors, trends, and inferences. It's about surfacing hidden insight that can help enable companies to make smarter business decisions.



**DATA MANAGEMENT**

Data management is the practice of organizing and maintaining data processes to meet ongoing information lifecycle needs. Emphasis on data management began with the electronics era of data processing, but data management methods have roots in accounting, statistics, logistical planning and other disciplines that predate the emergence of corporate computing in the mid-20th century.



There are mainly three kinds of Data

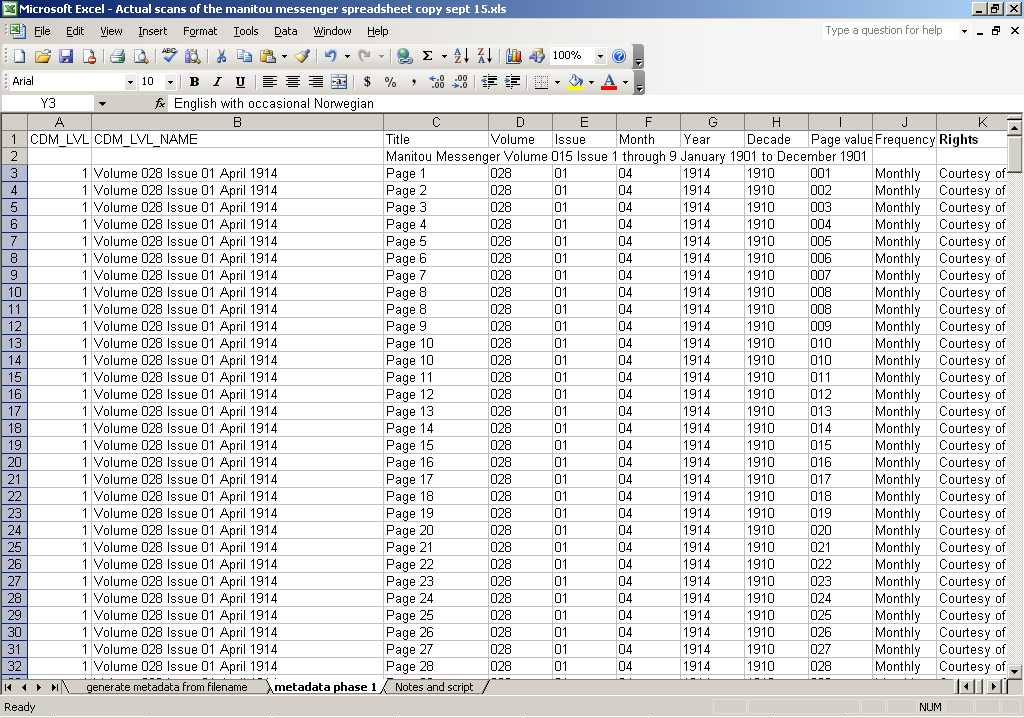
1) Structured Data

2) Unstructured Data

3) Semi Structured Data

**1)Structured Data:**

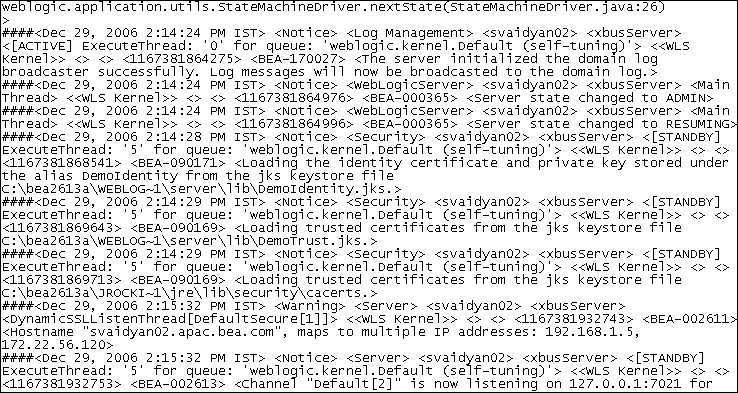
Structure Data is data that has been organized into aformatted repository, typically a database, so that itselements can be made addressable for more effective processing and analysis.



**2)Unstructured Data:**

The phrase Unstructured Data usually refers toinformation that doesn’t reside in a traditional row-column

database.Unstructured Data files often include text and multimediacontent.Examples include e-mail messages, word processing, documents, videos, photos, audio files, presentations, textcorpus etc.



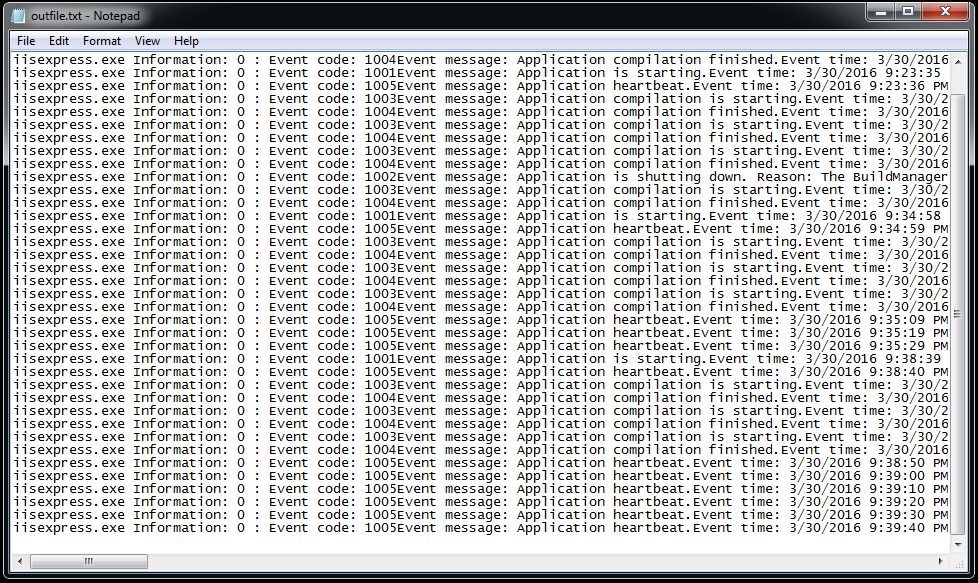
**3)Semi Structured Data:**

Semi-structured data is data that is neither raw data, nor

typed data in a conventional database system. Eg-CSV

(comma Separated Value), XML, JSON (JavaScript Object

Notation).



**Data Management Tools**

Good data management enables the location, sharing, and reuse of data, and reduces the redundancy of data. These attributes of good data management reduce costs in terms of time and money. Data management tools and best practices can also help to meet the open data requirements of making data more accessible and discoverable.

* When there is a structured and reasonable amount ofdata, the data tools are used to manage that data arecommonly known as database.

Like *Oracle*, *MySQL, Sql, CSV/Excel*.



* When there is an unstructured/ semi-structured andhuge amount of data or Big Data, the tool popularly used isknown as *Hadoop.*



**DATA ANALYTICS**

****

We have already learned about data previously. Now what do we mean by analysis? Analysis is detailed examination of the elements or structure of something what is data here. And now waht is idea? A thought or suggestion as to a possible course of action or the aim or purpose is somthing we call idea.

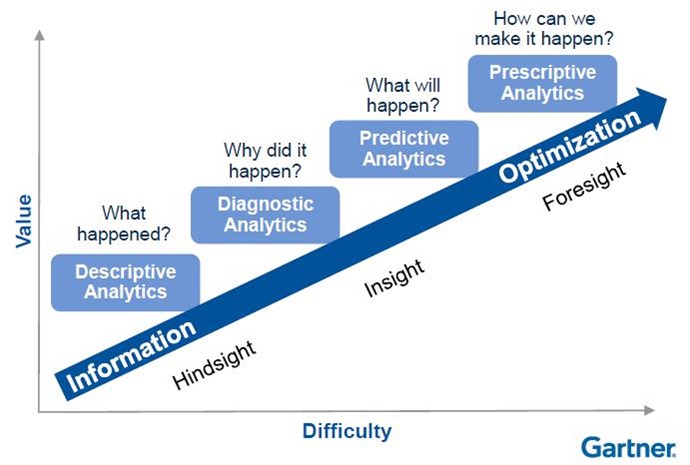
Now what is data analysis? Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making .

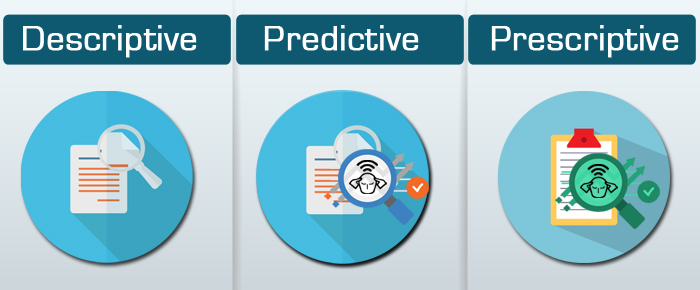
Now what are the pillars of data analytics, we are discussing that with a diagram below-



There are four types of data analytics,

1. Descriptive Analytics;
2. Diagnsostic Analytics;
3. Predictive Analytics;
4. Prescriptive Analytics.



* **Descriptive Analytics:**

Descriptive analytics is a preliminary stage of data processing that creates a summary of historical data to yield useful information and possibly prepare the data for further analysis.

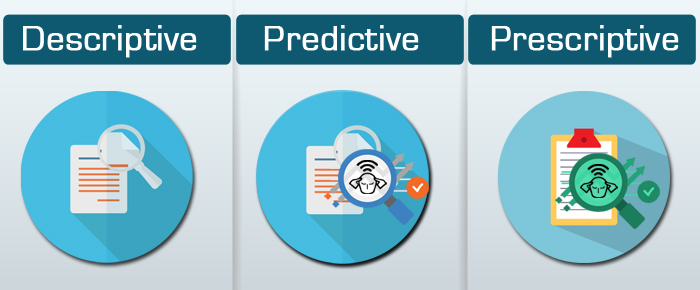
Descriptive analytics is sometimes said to provide information about happened.

* **Diagnsostic Analytics:**

 Diagnostic Analytics is a form of advanced analytics which examines data or content to answer the question “Why did it happen?”, and is characterized by techniques such as drill-down, data discovery, data mining and correlations.

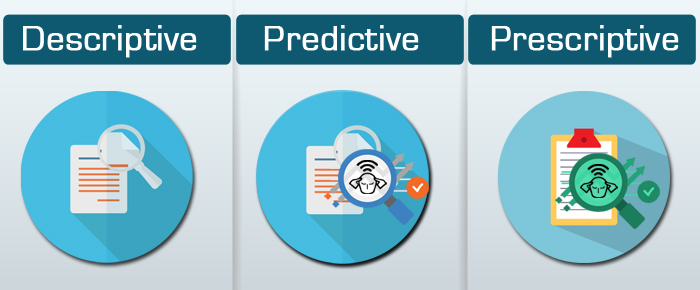
Diagnostic analytics is a deeper look at data to attempt to understand the causes of events and behaviors.

* **Predictive Analytics:**

 Predictive analytics is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends.

Predictive analytics, which is used to identify future probabilities and trends, is said provide information about what might happen in the future.

* **Prescriptive analytics:**

 Prescriptive analytics is the area of business analytics (BA) dedicated to finding the best course of action for a given situation. Prescriptive analytics is related to both descriptive and predictive analytics.

Prescriptive analytics is applied to try to identify the best outcome to events, given the parameters, and suggest decision options to best take advantage of a future opportunity or mitigate a future risk.

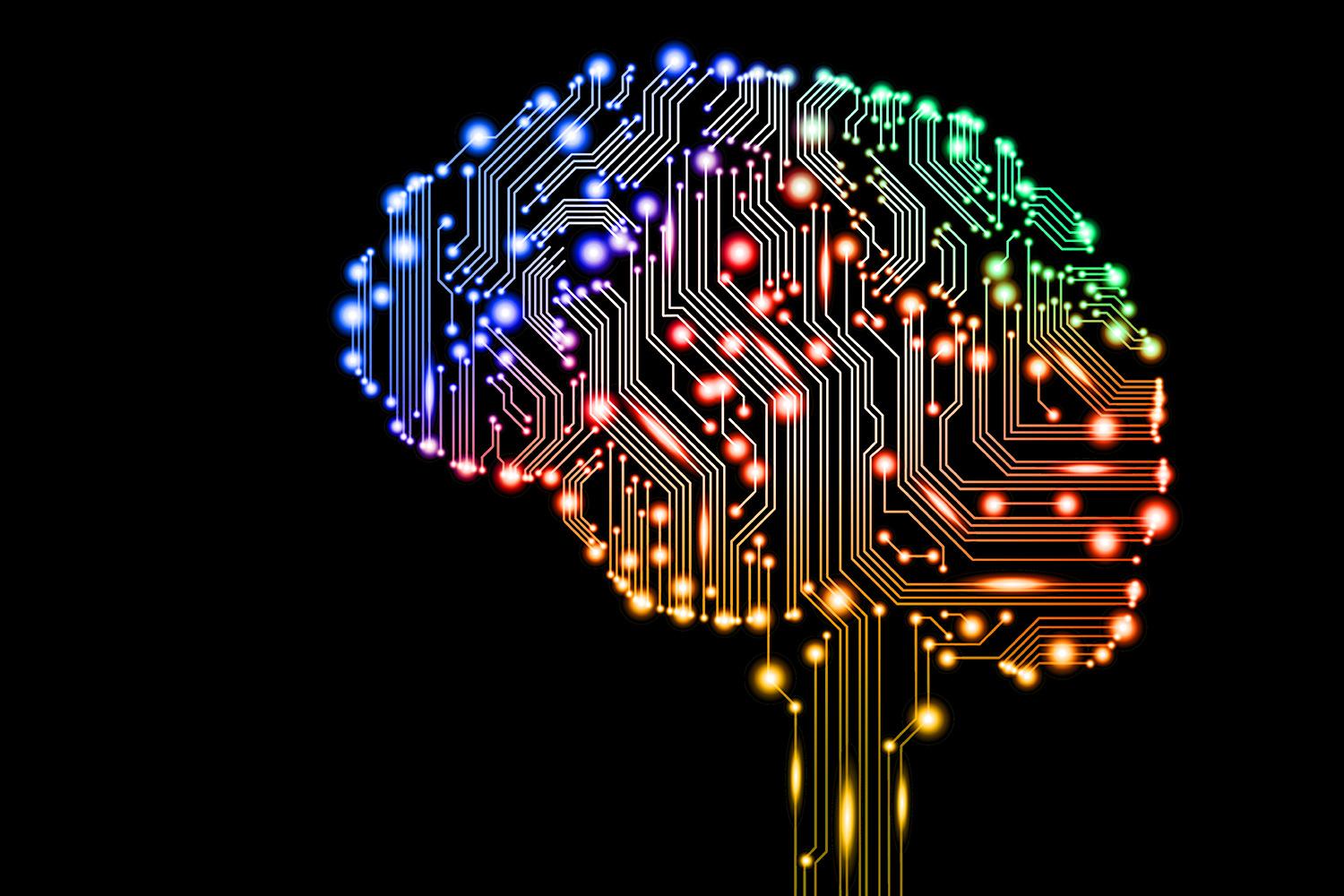
**Data Pre-processing**

**Data processing occurs when data is collected and translated into usable information.** Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly as not to negatively affect the end product or data output.

Data processing starts with data in its raw form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and utilized by employees throughout an organization.

**Artificial intelligence**

Artificial intelligence (AI) makes it possible for machines to learn from experience, adjust to new inputs and perform human-like tasks. Most AI examples that you hear about today – from chess-playing computers to self-driving cars – rely heavily on deep learning and natural language processing. Using these technologies, computers can be trained to accomplish specific tasks by processing large amounts of data and recognizing patterns in the data.



**Where we are implementing the concept of data science in HRData Analysis?**

Data analysis *discovering useful information, informing conclusions, and supporting decision-making* is a process of *inspecting, cleansing, transforming, and modeling data* with the goal of. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains. *In today's business, data analysis is playing a role in making decisions more scientific and helping the business achieve effective operation.*

Human Resourcedata is such a big data with varse number of entries like *15,000 employee observations and 10 features, and* HR analytics demonstrates the causal relationship between the activities exacted by an HR department and the business outcomes that result from this activity. Once this causal relationship is discovered, HR analytics professionals will devise and implement a strategic plan to assist the business in attaining better outcomes, to reach their goal.

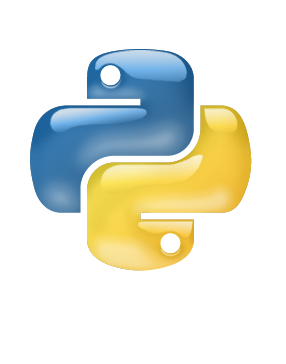
So, this is how we implement Data Science in our HR Analytics domain.

**Dataset**

The Data Set here used, was provided to us by our Mentor Tania Chakraborty of the NIVT training institute. The data is primarily a raw data of human resource records, we analyse data using python.

This data set is said to be original data entries from very known American multinational information technology company, International Business Machines Corporation (IBM). Who have dealt operations in over 170 countries.

In this data analysis report we are *using python as our data analysis tool.* Because, python is a high-level programming language designed to be easy to read and simple to implement. It is open source, which means it is free to use, even for commercial applications. Python can run on Mac, Windows, and Unix systems and has also been ported to Java and .NET virtual machines.



Now to implement the codes needed for analysis we need a platform that is *Jupyter Notebook* here. *The Jupyter Notebook is an open-source web application* that allows you to create and share documents that contain live code, equations, visualizations and narrative text. . Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.



* The Setting:

Now let’s say IBM has hired you as a data scientist to help them make a business decision.

Here’s the question they want answered:

* Find out the Data Description.
* Find out the Statistical Description of the data.
* Find out the correlation matrix of all the variable present in the dataset and visualize the matrix.
* Plot counts the number of employee who left the company and who didn’t.
* Count plot to check the number of employees with respect to Salary and Department feature and grouping together
* Plot a diagram to check the number of employees left with respect to Salary and Department feature and grouping together.
* How we can predict that an employee’s is salary with respect to satisfaction level, years at company, project count.

To help you make this decision, the company has provided you with a .CSV file that contains statistics on how employees are behaving in such manner.



“The goal is to turn data into Information, and information into Insight.”

-Carley Fiorina.

(CEO of HP from 1999 to 2005)

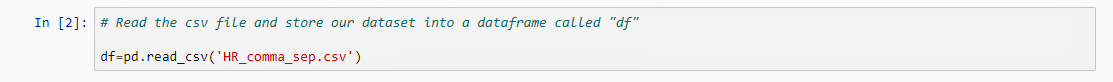
Code & Comments

Obtaining the Data

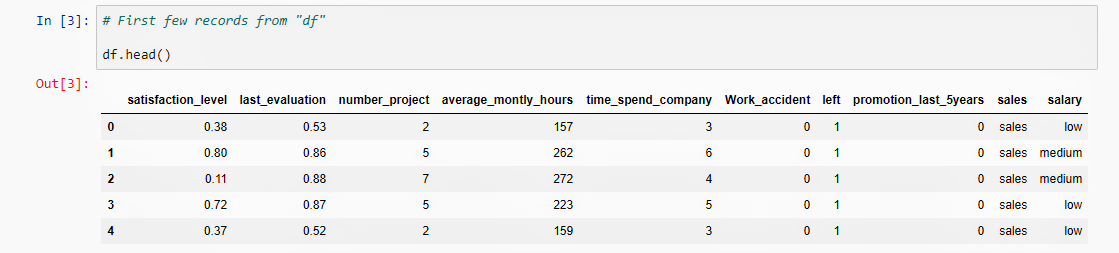


*Firstly, let’s import the Python libraries we will use for data preparation and analysis on Jupyter Notepad.*

*We have imported* pandas *for linear algebra,* numpy *for data processing,* matplotlib.pylot *for data visualization,* seaborn *for statistical plotting. And to ignore the warnings we have imported* warnings *and then set those warnings to ‘ignore’.*

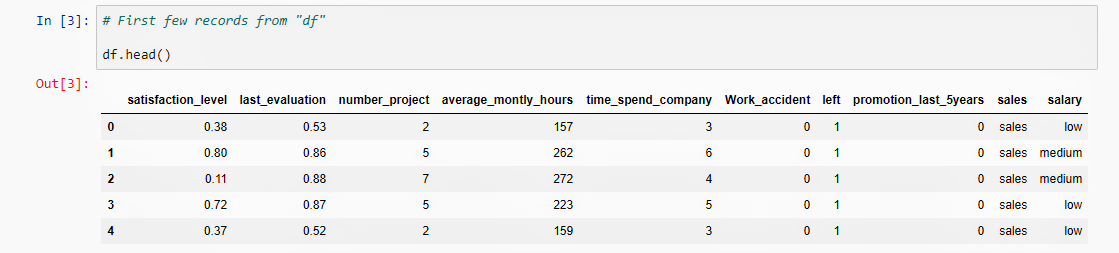
**

*Then, we can import the CSV file ‘*HR\_comma\_sep*’ using the*.read\_csv ()*method from pandas and store it in a data frame called ‘*df *’.*

**

*Here we are checking the first few record from ‘df ’ where we have stored the dataset, by using the* .head() *method.*

*And the output of* In[3] *is –*

**

Scrubbing the Data

**

*Now we are checking if there are any null or missing values in the dataset or not using the method* .isnull().any()*.*

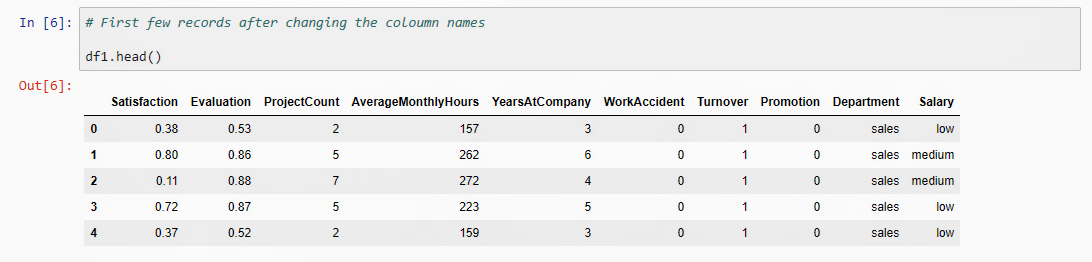
*And the output of* In[4] *is –*

**

*So, we are seeing that there are no missing data here.*

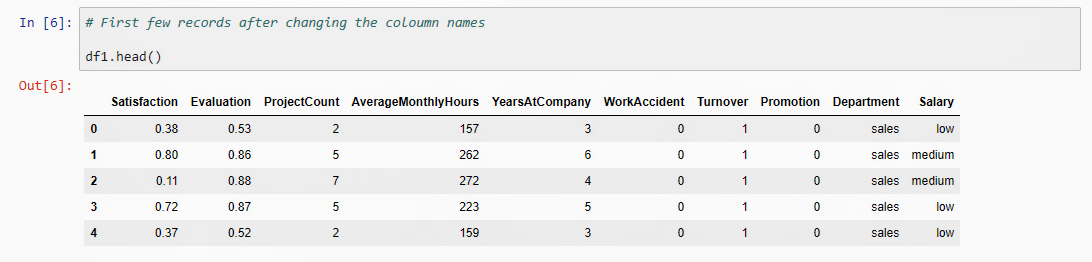
**

*In this line we have renamed certain columns for better readability to understand this dataset in a better way by using* .rename *method and stored in into new data frame named* ‘df1’*. Now the new column names are* ‘Satisfaction’, ‘Evaluation’, *‘*ProjectCount’, ‘AverageMonthlyHours’, ‘YearsAtCompany’, ‘WorkAccident’, ‘Turnover’, ‘Promotion’, ‘Department’, ‘Salary’ *respectively.*

**

*Again we are checking the first few records of the data set after renaming the column names.*

*And the output of* In[6] *is –*

**

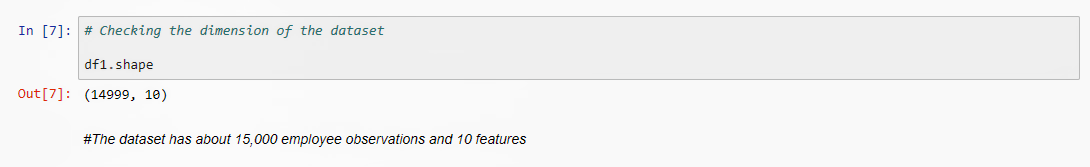
Exploring the Data

“Torture the Data, and it will confess evrything”

-Ronald Coase,

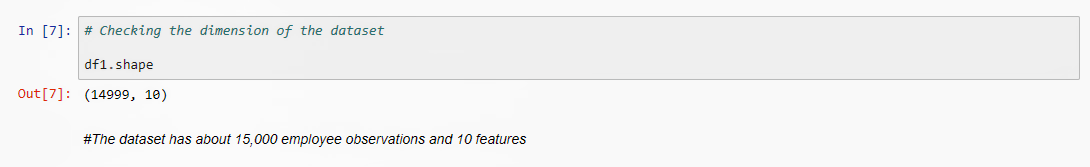
(Who believed economists should study real markets and not theoretical ones, established the case for the corporation as a means to pay the costs of operating a marketplace. *He received the Nobel Prize in Economics in 1991*.)

Statistical overview

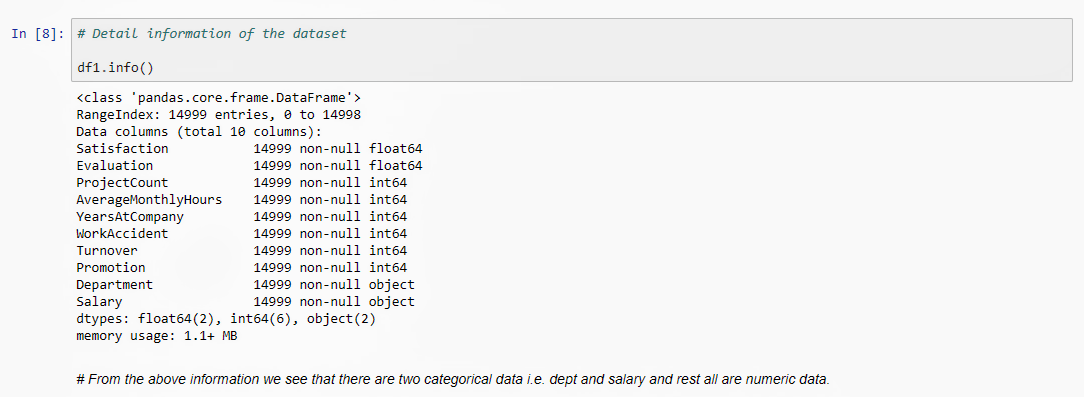
**

*Now we are checking the dimensions of the dataset by using the method* .shape.

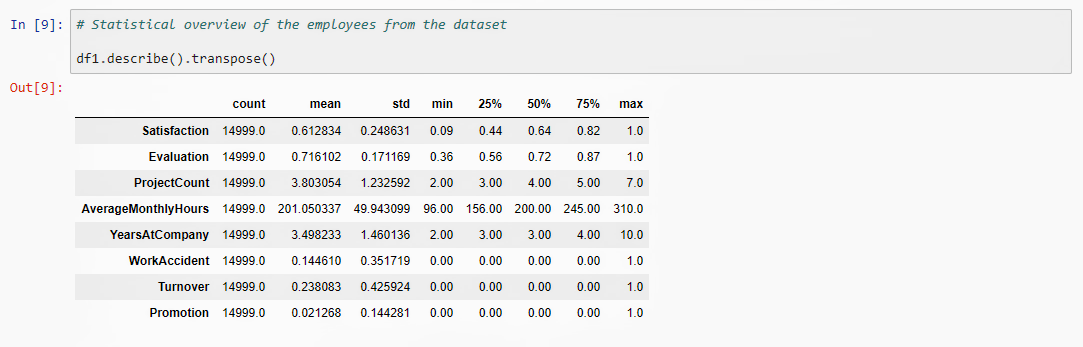
*And the output of* In[7] *is –*

**

*So, the dataset has 15,000 columns which is the employee observations and 10 rows which is features of* ‘df1’.

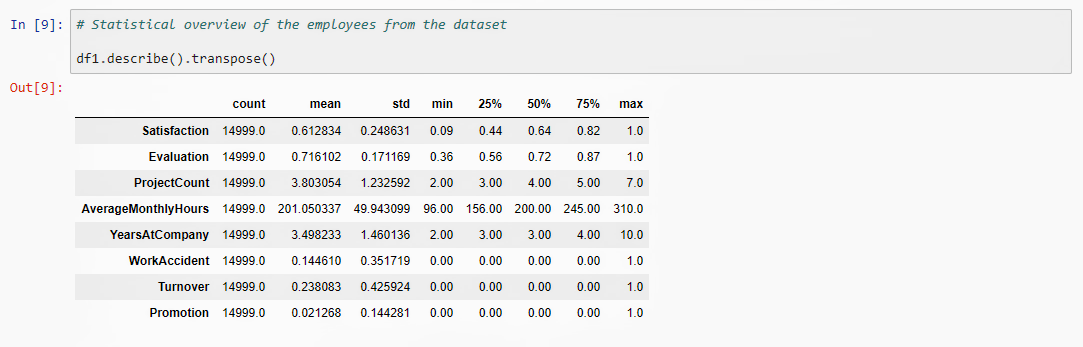


*Now from this line we are getting the detail information of the dataset by using the method* ‘.info()’. *And we see that there are two categorical data that is ‘*Department’ *and ‘*Salary*’ and rest all are numeric data.*

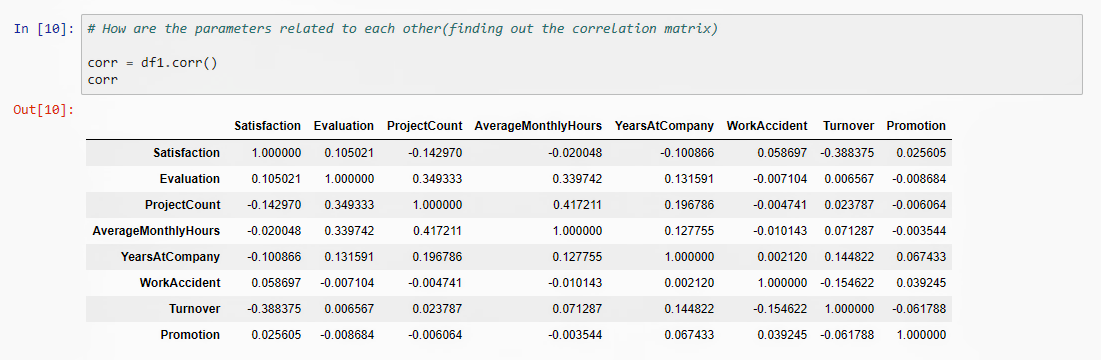
**

*By this command we are giving a look on statistical overview of the employees from the dataset by using* .describe() *method. And for a vertical mannered table we used* .transpose() *method.*

*And the output of* In[9] *is –*

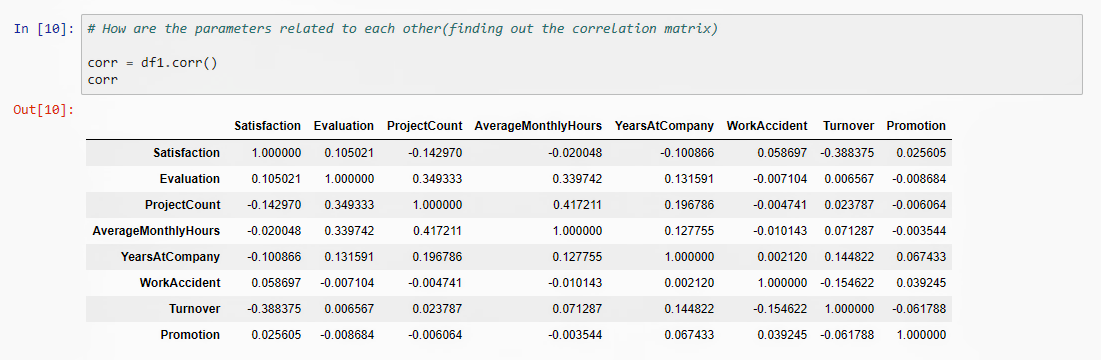
**

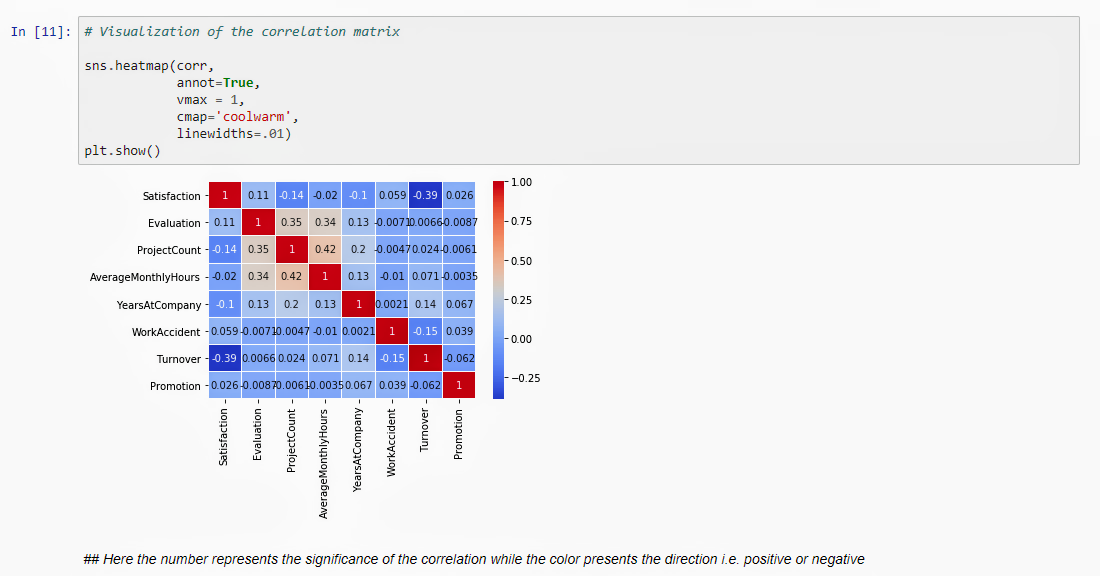
Correlation Matrix & Heat-Map

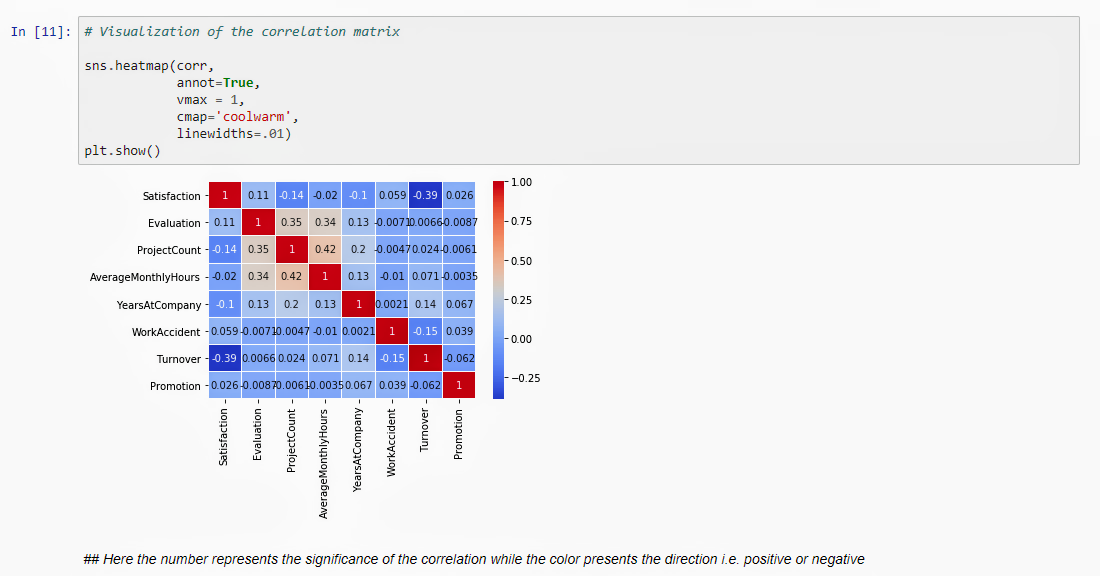
**

*There are many parameters in this dataset. To find a relation between them we are finding out correlation matrix by using* .corr() *method. And stored it into a variable, named ‘*corr*’.*

*And the output of* In[10] *is,*

**

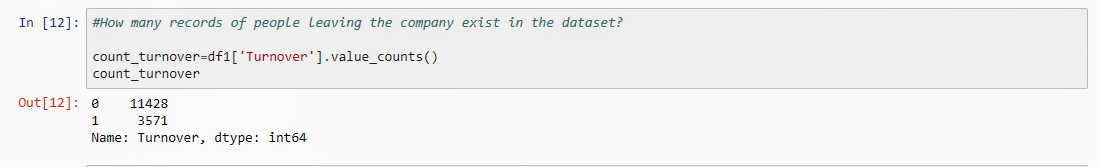
**

**

## Here the number represents the significance of the correlation while the colour presents the direction i.e. positive or negative.

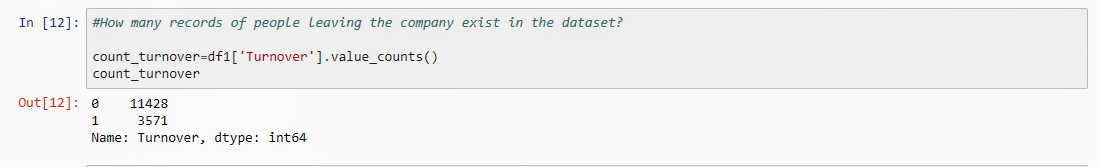
*Now from this line we visualized the correlation matrix by* seaborn.heatmap() *method where we have plotted the ‘*corr*’ variable. Here the blue colour represents the positive(pale blue) and negative(deep blue) directions.*

Turnover & Non-Turnover Analysis

**

*Here we are having a quick overview on the records of people leaving the company exist in the dataset by using the* .value\_counts() *method. And we are storing it into a variable names ‘*count\_turnover*’.*

*And the output of* In[12] *is,*

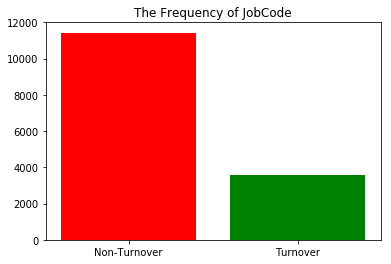
**

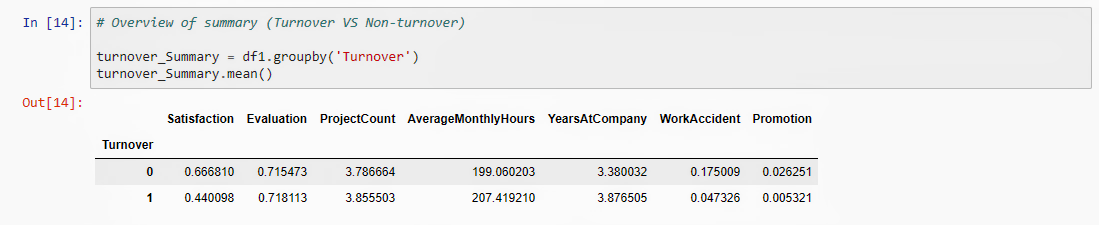
*Here ‘0’ is true for turnover value while ‘1’ is for false. So 11,428 people left the company while 3,571 people stayed out of 15,000 records of people here.*

**

*Then we visualized the records of people leaving the company with the help of a bar diagram using the method* matplotlib.pyplot.bar()*. And gave this diagram a title ‘The Frequency of JobCode’ by using* matplotlib.pyplot.title() *method.*

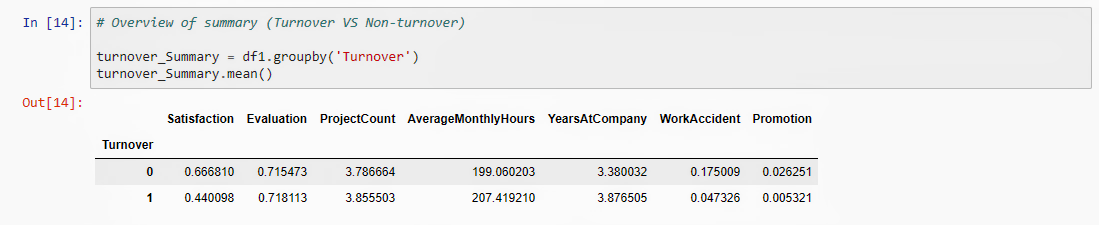
*And the output of* In[13] *is,*

**

**

*In this line we are giving a look on quick overview summary on the data of Turnover Vs. Non-Turnover employee. By using* .groupby() *method we are separating the all ‘Turnover’ data and grouping them into a variable named ‘*turnover\_summary’*. And then by using the* .mean() *method we summarized that variable data.*

*And the output of* In[14] *is,*

**

Salary distribution& Turnover Analysis

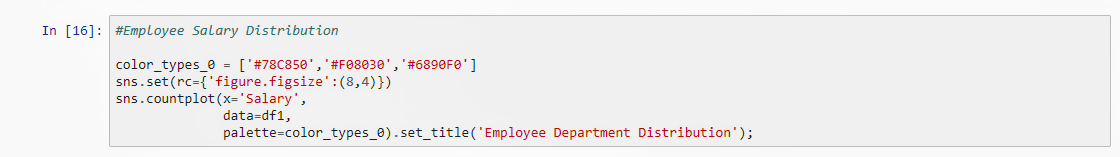
**

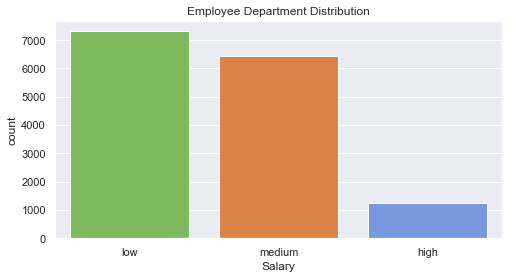
*Next we are checking the records of salary types, means how many unique categorical values are there in the column* ‘Salary’ *by using the* .unique() *method.*

*And the output of* In[15] *is,*

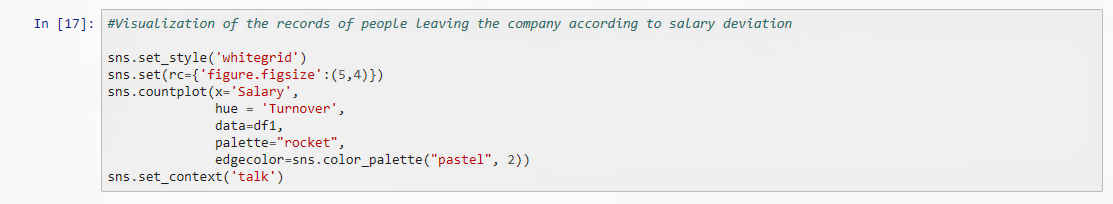
**

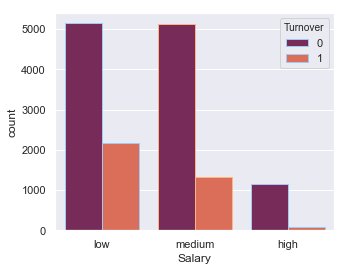
*So, there are three unique data types in ‘Salary’ column,which are* ‘low’, ’medium’, ‘high’.

**

**

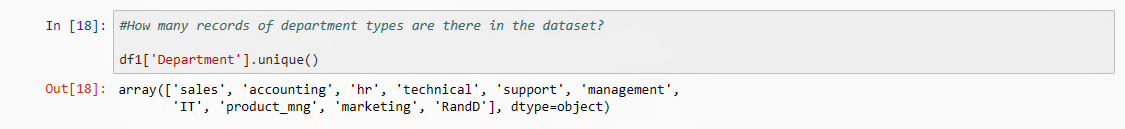
*Here in this line we are visualizing employee salary distribution with countplot diagram from the dataset by using the method* seaborn.countplot()*. And the diagram is titled with* seaborn.title() *method as ‘*Employee Salary Distribution*’.*

**

**

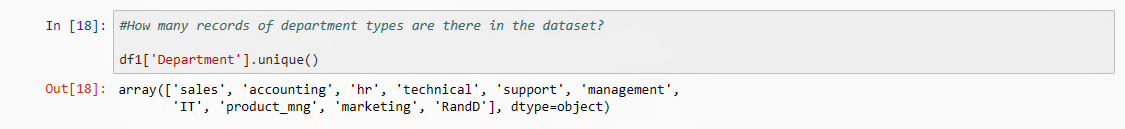
*Now we are visualizing the records of people leaving the company according to salary deviation with the countplot diagram from our dataset by using the method* seaborn.countplot()*. Here we have plotted the Salary deviation Vs. Turnover values respectively.*

Department distribution & Turnover Analysis

**

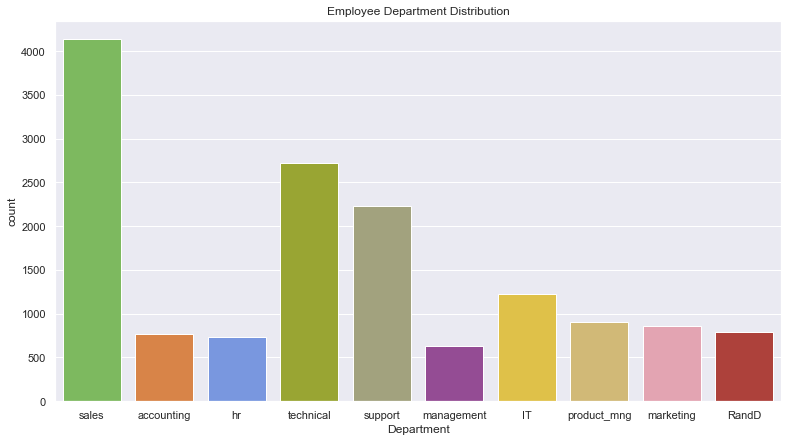
*Here we are checking the records of department types, means how many unique categorical values are there in the column* ‘Department’ *by using the* .unique() *method.*

*And the output of* In[18] *is,*

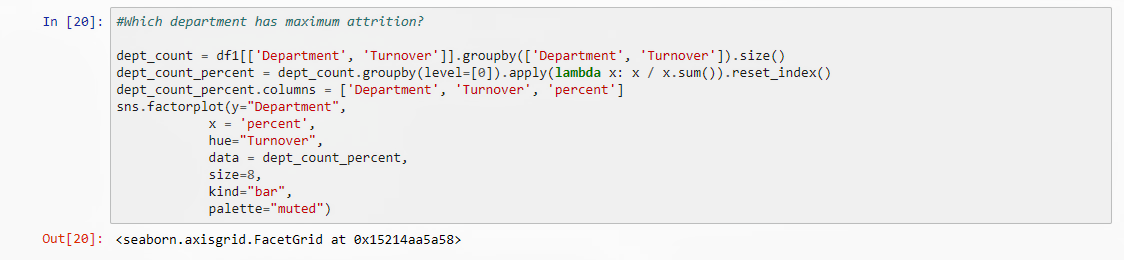
**

*So, there are ten unique data types in ‘*Department*’ column,which are* 'sales', 'accounting', 'hr', 'technical', 'support', 'management', 'IT', 'product\_mng', 'marketing', ‘RandD’.

**

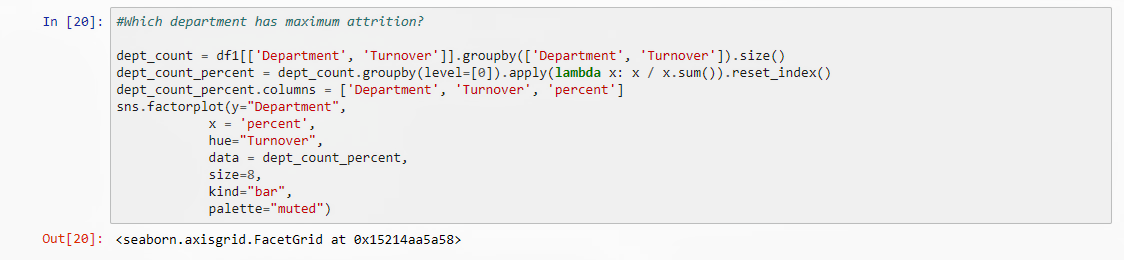
**

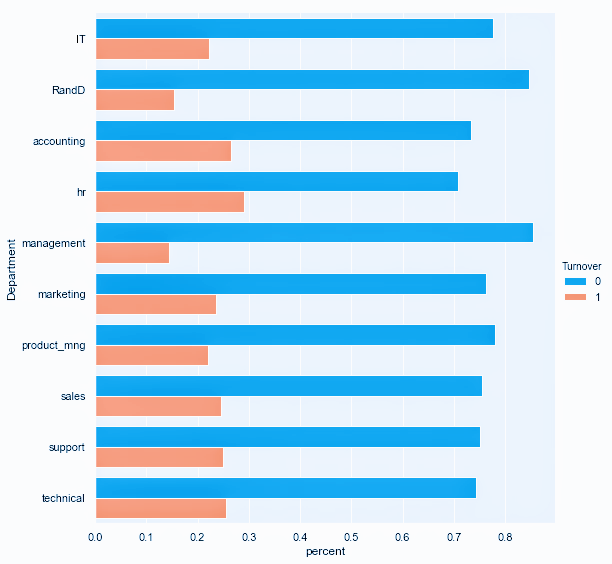
*Here in this line we are visualizing employee department distribution with countplot diagram from the dataset by using the method* seaborn.countplot()*. And the diagram is titled with* seaborn.title() *method as ‘*Employee Department Distribution*’.*

**

*To visualize the Department Vs. Turnover we have to group the department and turnover values by using* .groupby()*method in a variable named* ‘dept\_count’*. Then we calculate the percentage according to the sum of each row from the* ‘dept\_count’*with the help of ‘*lambda function’ *and stored into a new variable named ‘*dept\_count\_percent’ *and, creating a new column* ‘percent’. *Using the method* seaborn.factorplot() *we visualize that which department has maximum attrition with a factorplot diagram.*

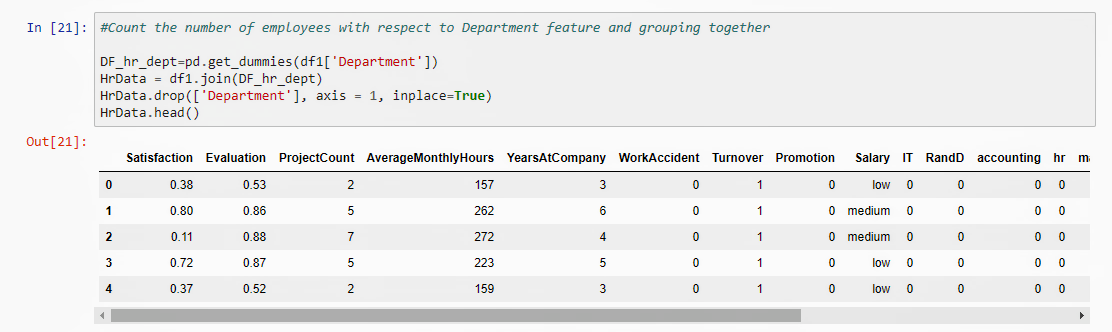
*And the output of* In[20] *is,*

**

**

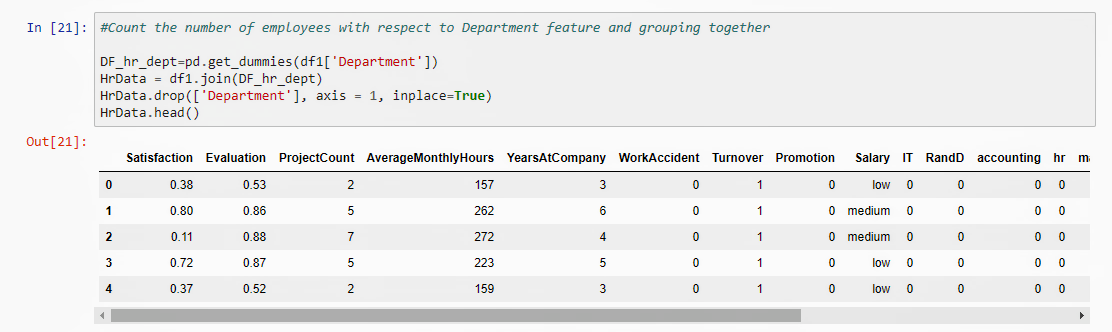
*Now we have a clear view of which department had the maximum turnover and the non-turnover as well.*

Department Vs Salary

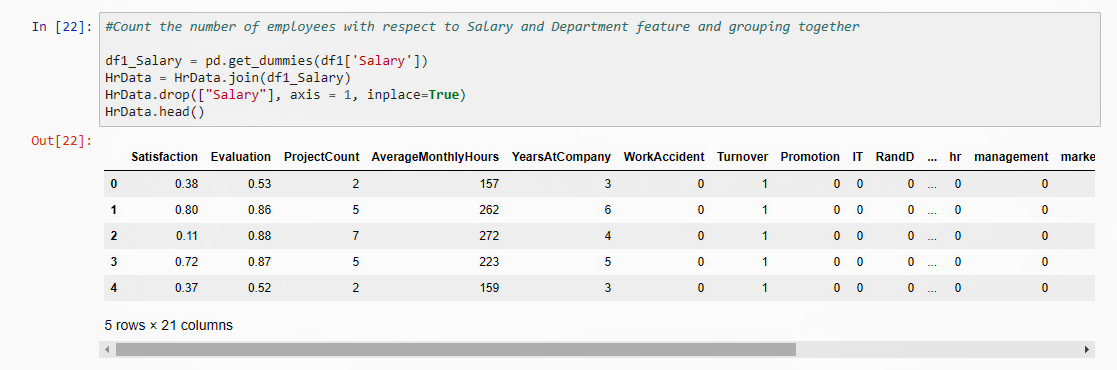
**

*As we need to count the number of employees with respect to department feature and grouping together, we have to create a dummy variable of department column by using* pandas.get\_dummies()*method and store it into a new variable named* DF\_hr\_dept. *Again in a new variable named* HrData *we joined the* DF\_hr\_dept *with* df1*dataframe and dropped the old* Department *column by using* .drop() *method. Hence checking first few records using* .head() *method.*

*And the output of* In[21] *is,*

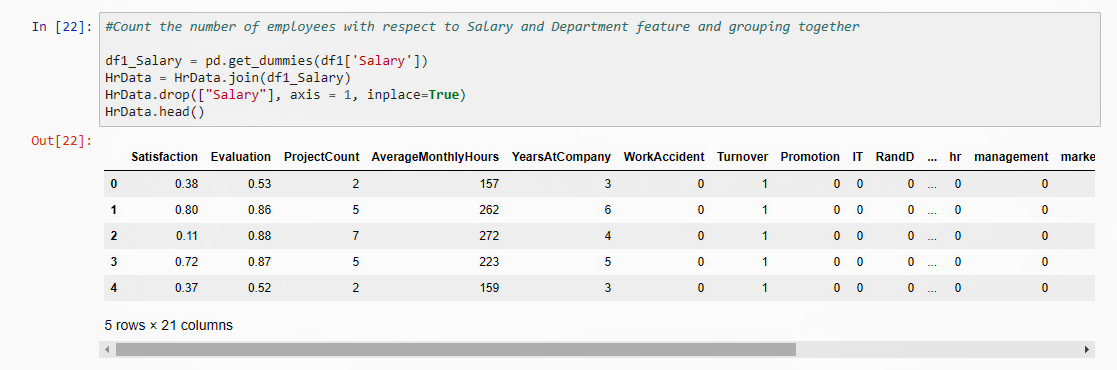
**

*So here we are noticing that this data frame does not containing the Department column instead of having ten unique departments named columns.*

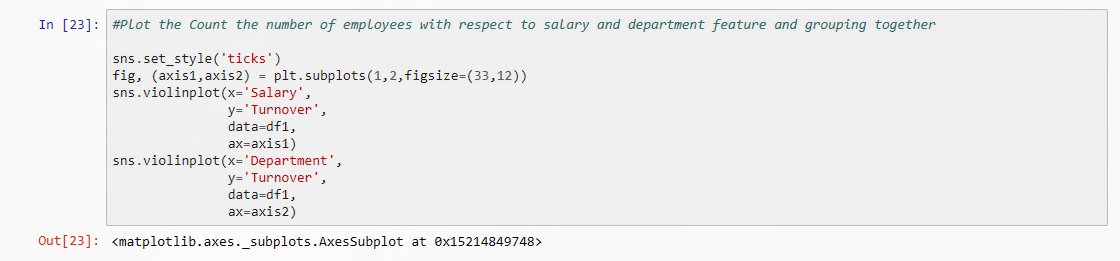
**

*As we need to count the number of employees with respect to Salary and Department feature and grouping together, we have to create a dummy variable of Salary column by using* pandas.get\_dummies()*method and store it into a new variable named* df1\_Salary. *Again in* HrData *we joined the* df1\_Salary *with very last updated* HrData *and dropped the old* Salary *column by using* .drop() *method. Hence checking first few records using* .head() *method.*

*And the output of* In[22] *is,*

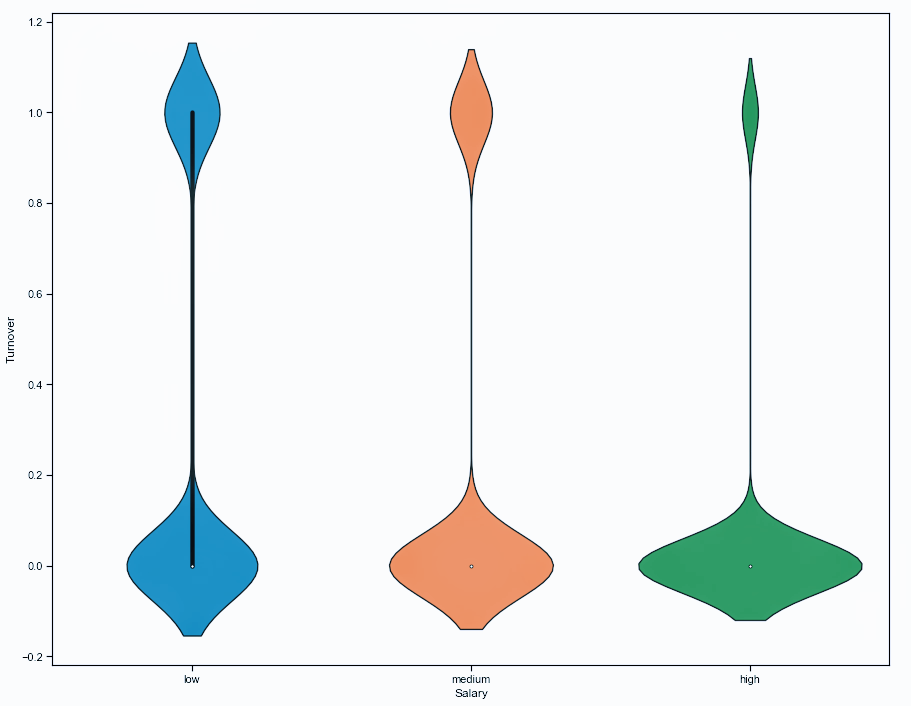
**

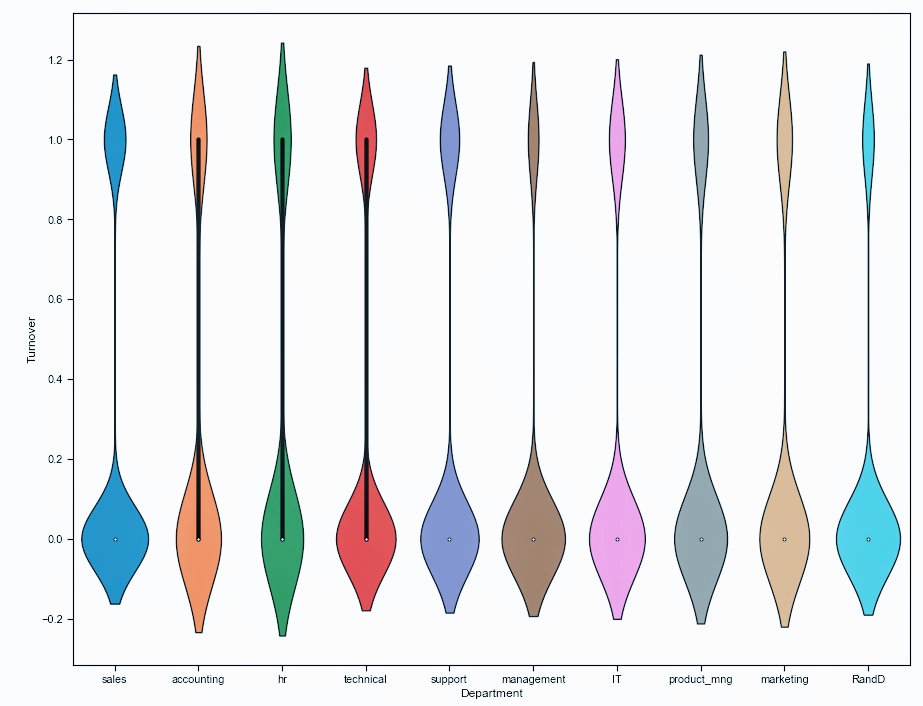
*So here we are noticing that this data frame does not containing the Department and Salary column instead of having ten unique departments and three unique salary named columns.*

**

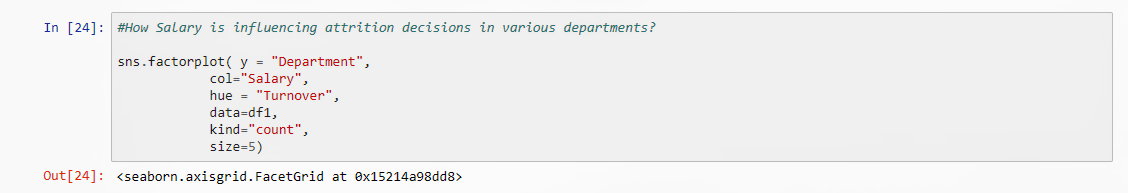
*Now to visualize the frequency of employees with respect to grouped salary and department feature together in terms of turnover and non-turnover we are going to plot a violin diagram from that dataset by using* seaborn.violinplot() *method.*

*And the output of* In[23] *is,*

**

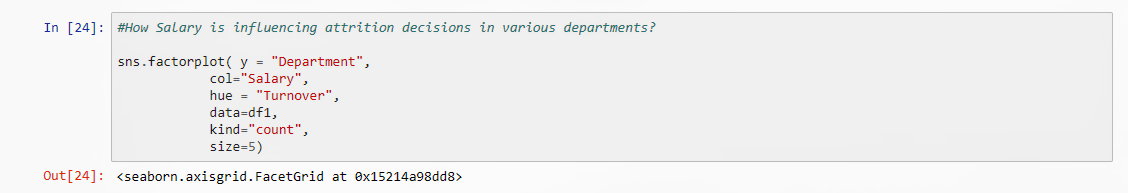
**

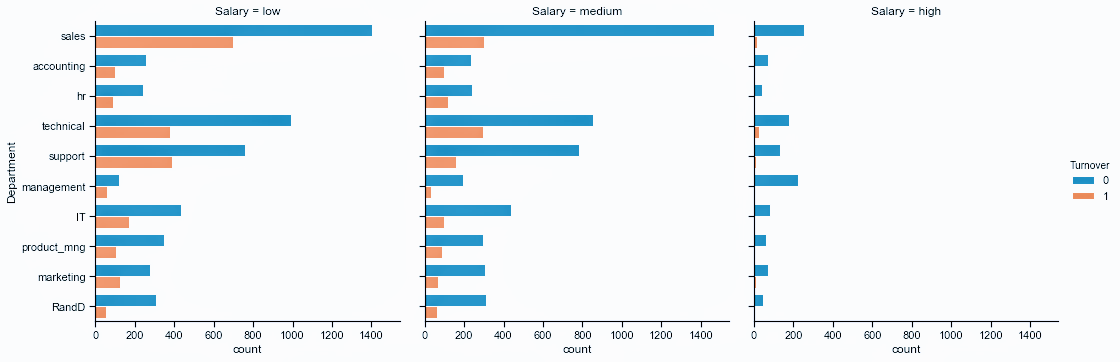
*Now we had a view on the frequency of employees with respect to grouped salary and department feature together in terms of turnover and non-turnover through this violin diagram.*

**

*Now we are visualizing how salary is influencing attrition decisions in various departments by using the method* seaborn.fatorplot()*with the help of factor plot. Here we are comparing the department wise turnover with respect to salary deviation.*

*And the output of* In[24] *is,*

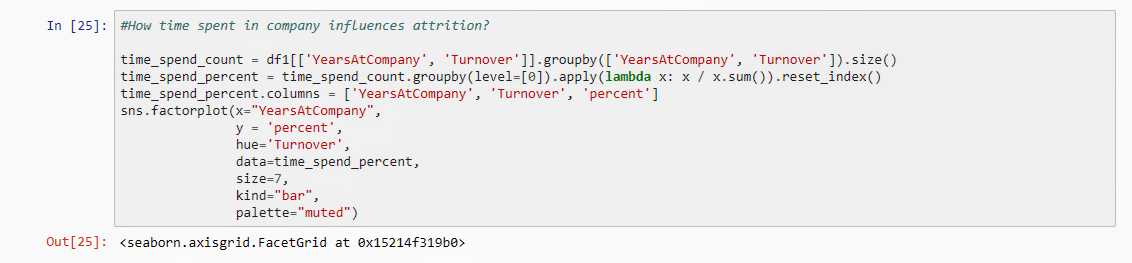
**

**

*Now we have a visualized factor plot diagram on salary deviation Vs department deviation.*

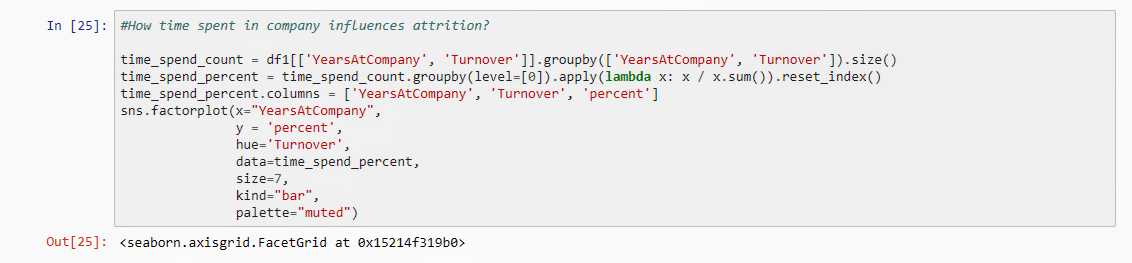
Time Spent Vs Turnover

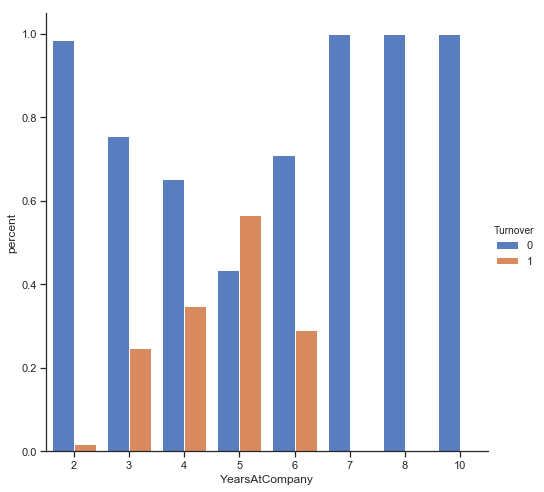


**

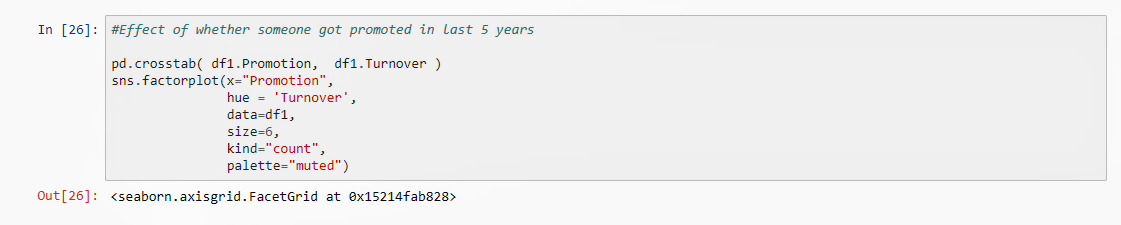
*Now we are visualizing how time spent in company is influencing attrition decisions by using the method* seaborn.fatorplot() *with the help of factor plot. So we grouped the years at company and turnover data in a new variable named ‘*time\_spend\_count*’ and then we counted the turnover values with the help of* lambda function *in a new variable again named ‘*time\_spend\_percent*’.*

*And the output of* In[25] *is,*

**

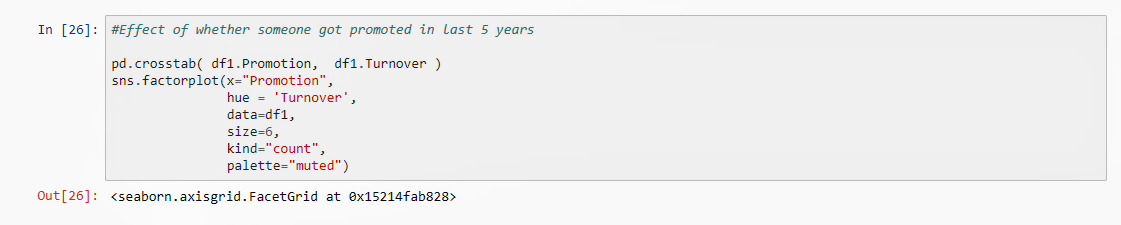
**

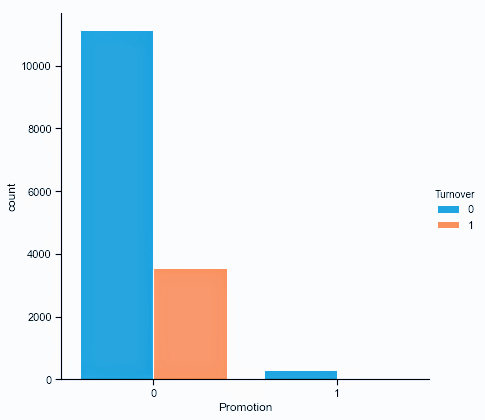
Promotion Vs Turnover

**

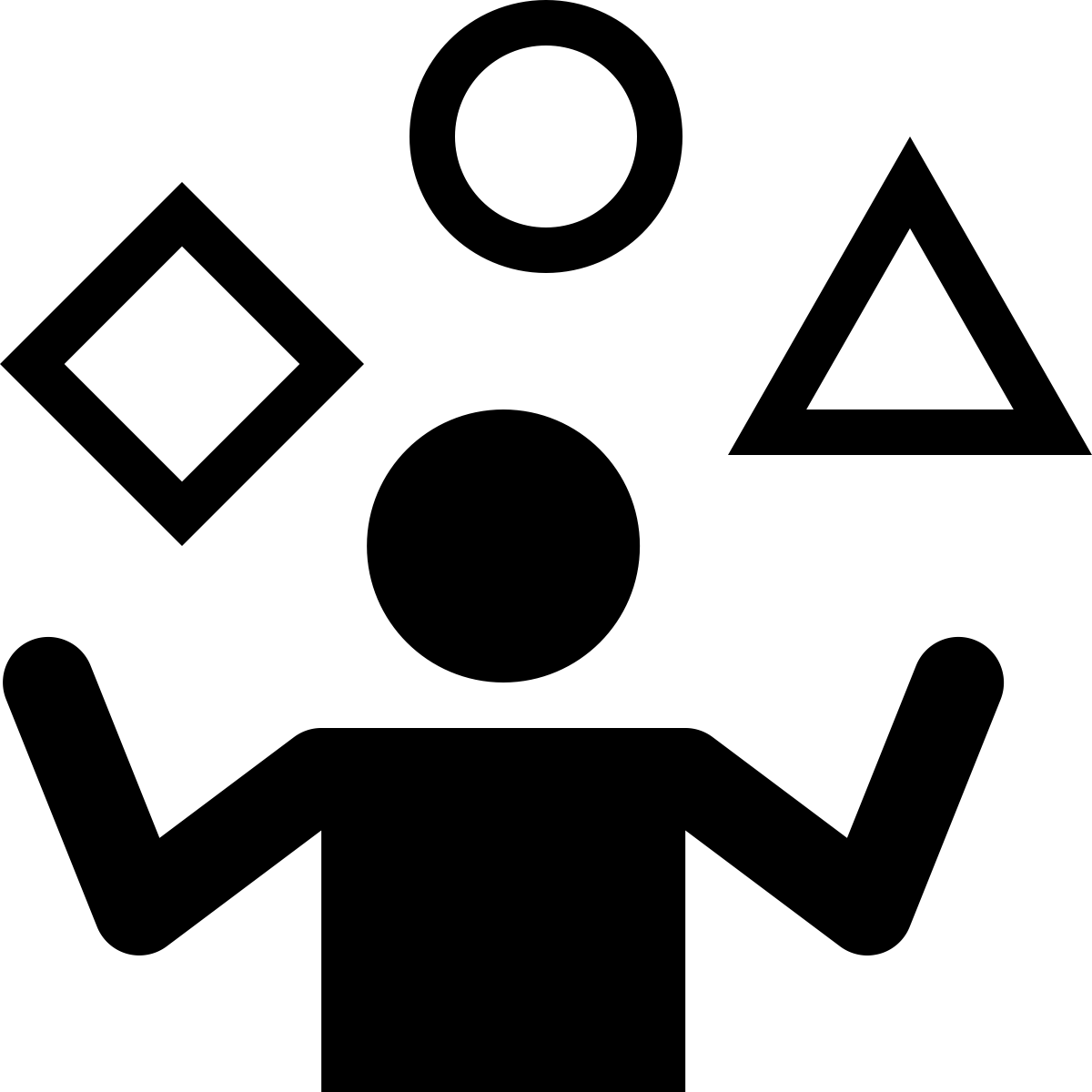
*Now we are visualizing how promotion in last 5 years is influencing attrition decisions by using the method* seaborn.fatorplot()*with the help of factor plot. Here we made a crosstab by using* pandas.crosstab() *method on promotion and turnover data from the given dataset.*

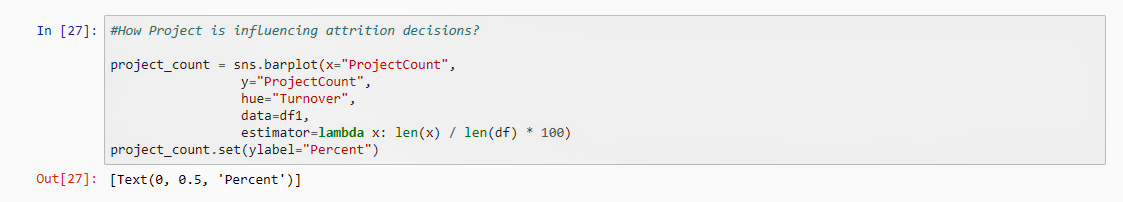
*And the output of* In[26] *is,*

**

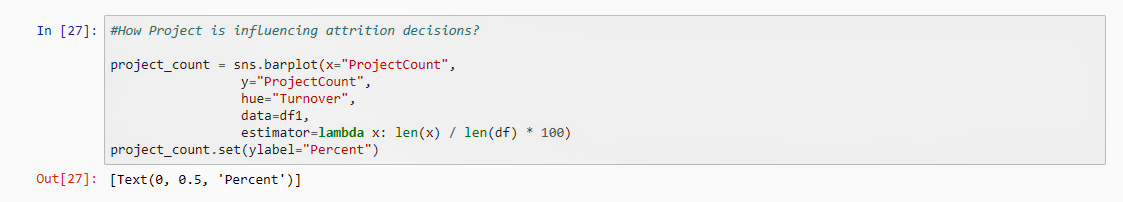
**

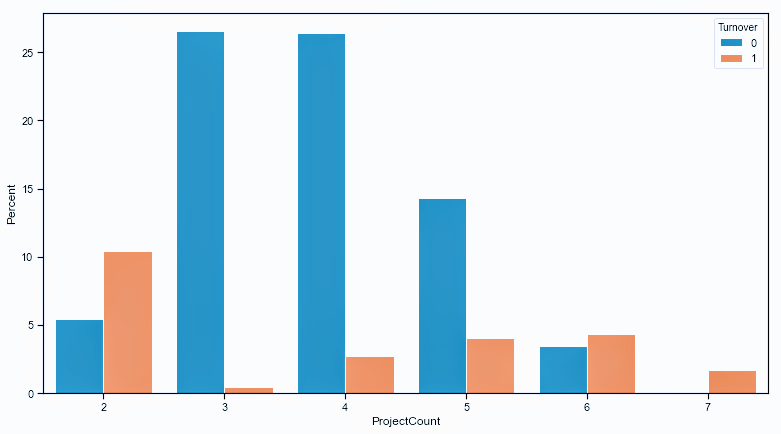
Project Done Vs Turnover

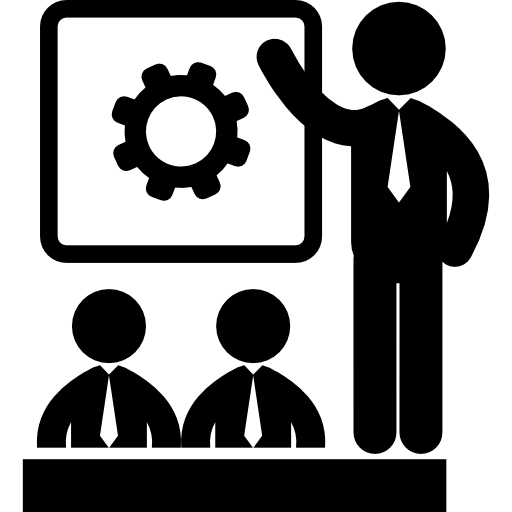
**

**

*Now we are visualizing how project is influencing attrition decisions by using the method* seaborn.barplot()*with the help of bar diagram. And the output of* In[27] *is,*

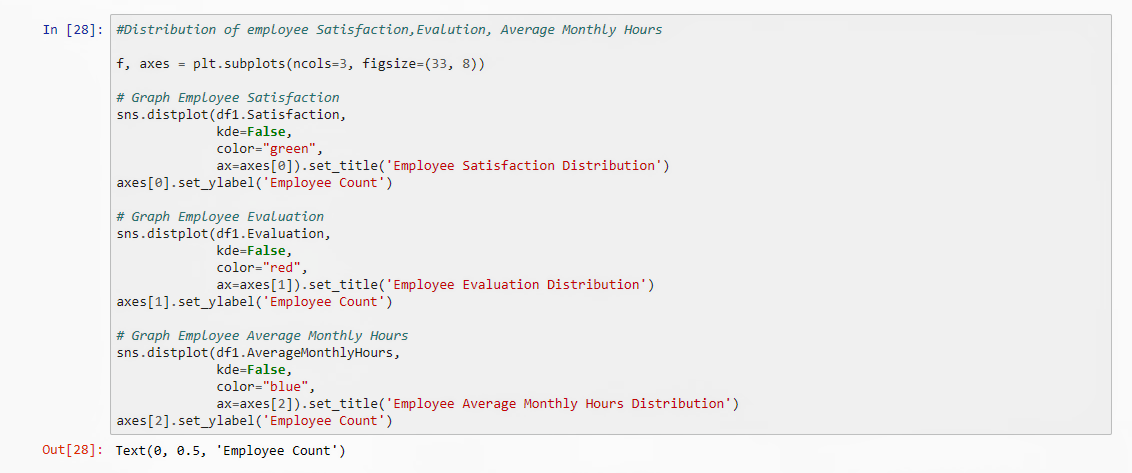
**

**

**

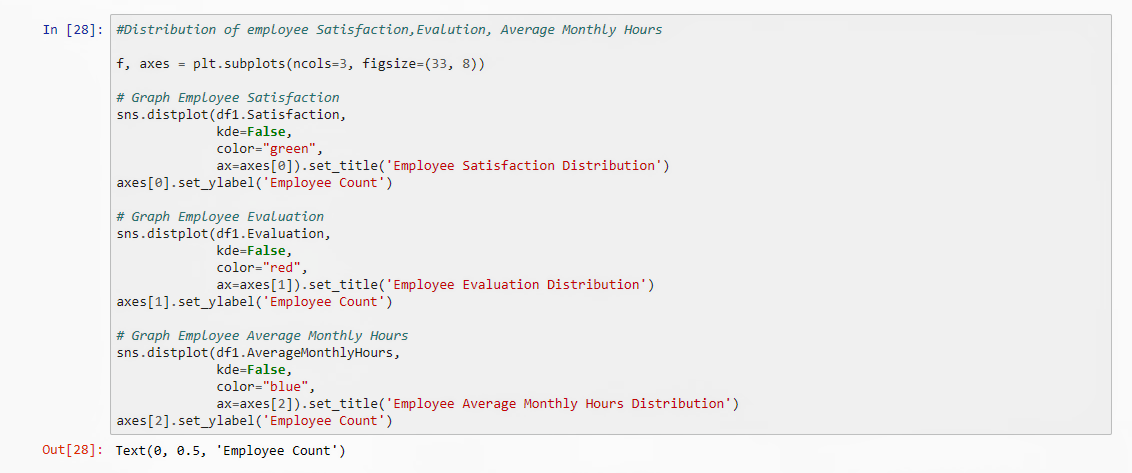
Distribution of Employee Satisfaction, Evaluation and Average Monthly Hours & Turnover Analysis

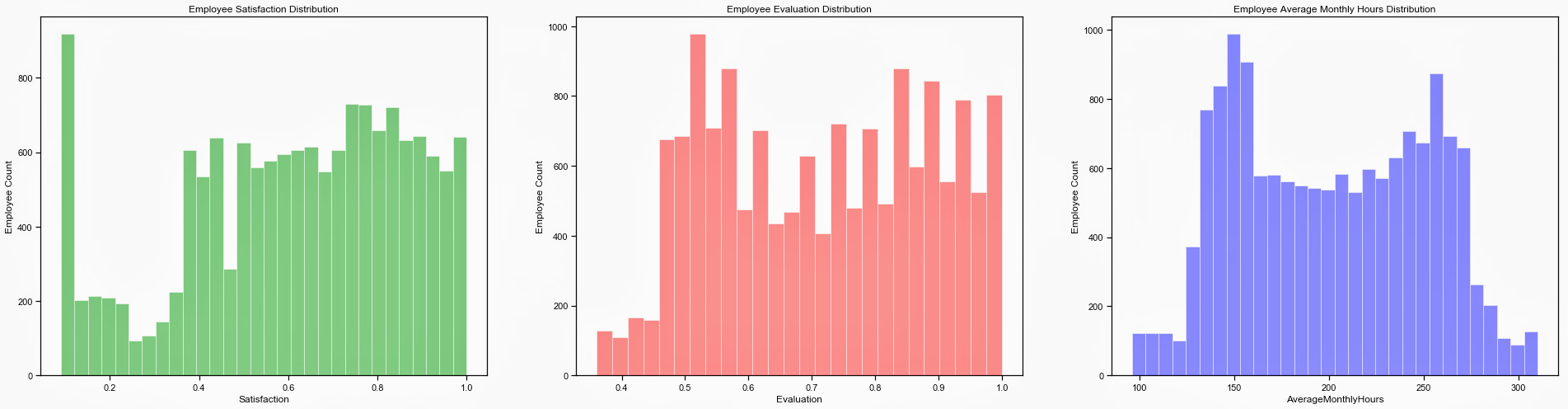
**

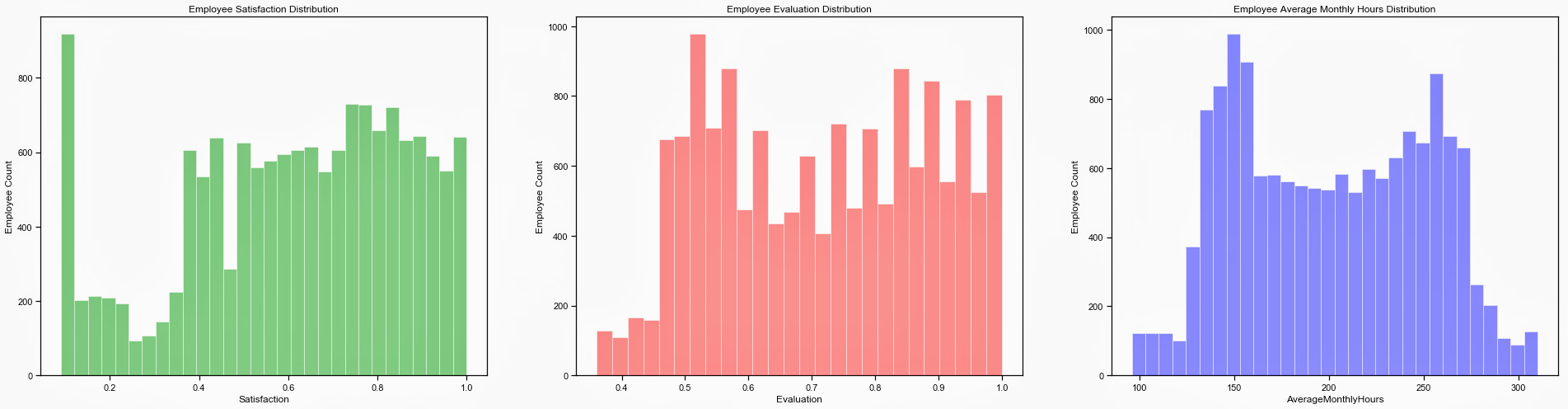
**

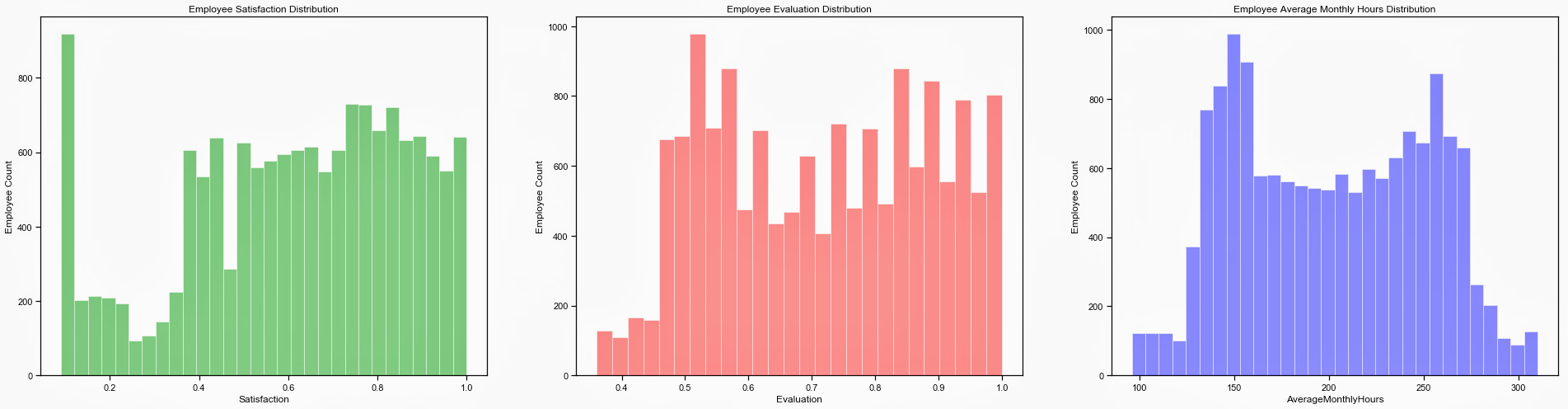
*Now we are visualizing the distribution of employee satisfaction, evaluation and average monthly hours frequencies by using the method* seaborn.distplot()*with the help dist diagram. And we made these three diagrams in a single plot by setting axes index.*

*And the output of* In[28] *is,*

**

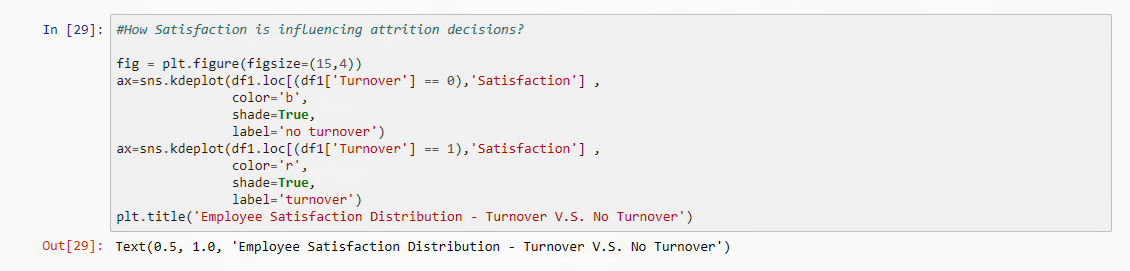
**

**

**

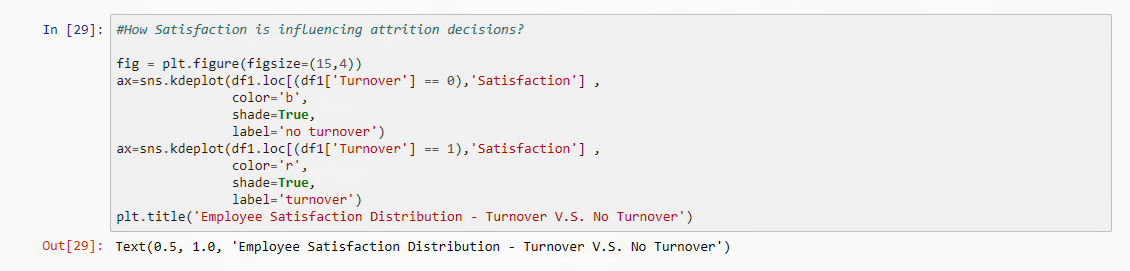
*Now we had a distribution of those three variable frequencies here.*

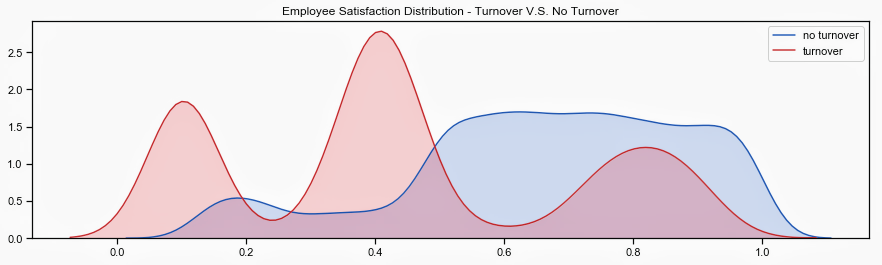
**

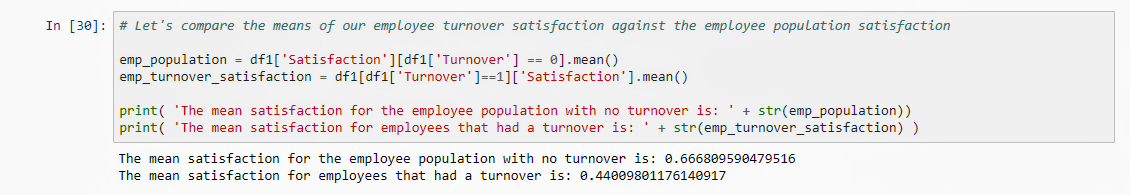
**

*Now we are visualizing how satisfaction is influencing attrition decisions by using the method* seaborn.kdeplot()*with the help of kernel density estimation(kde) diagram. Here we are comparing the satisfaction with respect to turnover and non-turnover. And gave a title ‘*Employee Satisfaction Distribution – Turnover V.S. No Turnover*’ by using* matplotlib.pyplot() *method.*

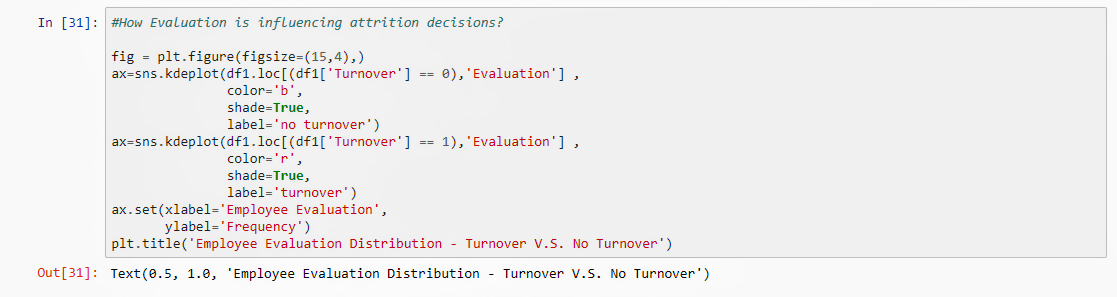
*And the output of* In[29] *is,*

**

**

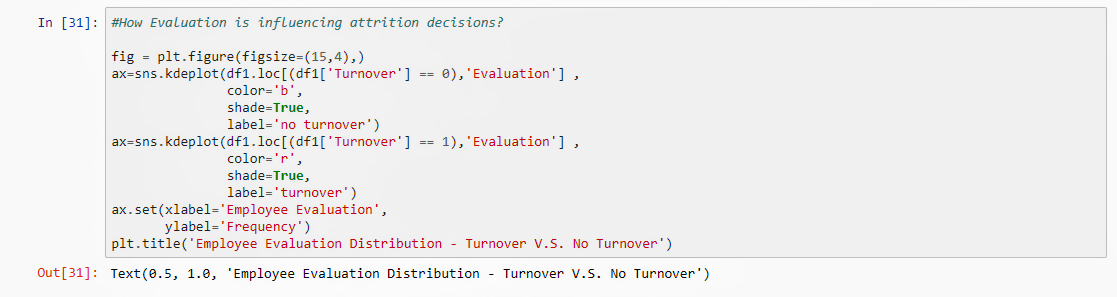
**

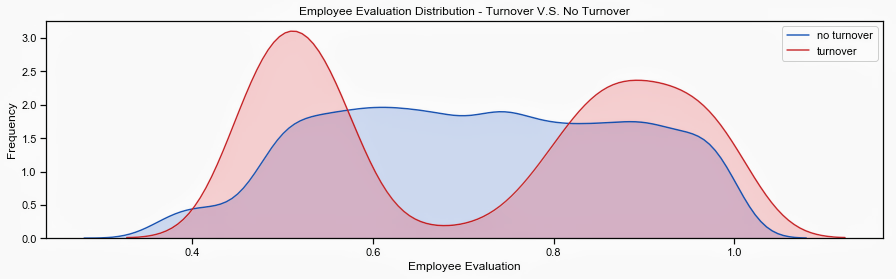
*Now in this line we have compared the mean employee turnover satisfaction with the satisfaction of the employee stayed in the company.*

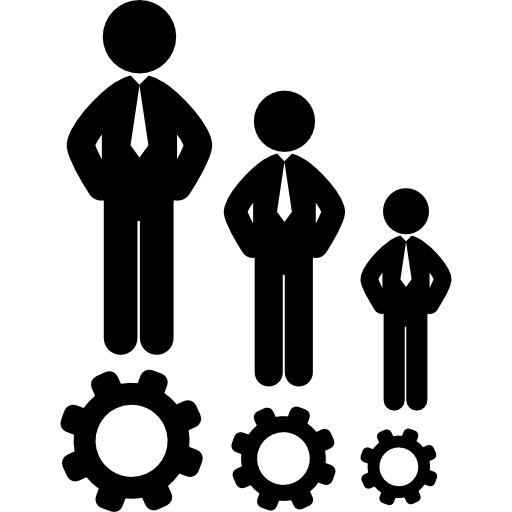
**

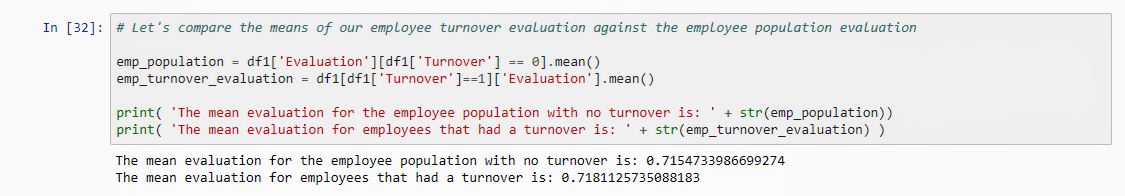
*Now we are visualizing how evaluation is influencing attrition decisions by using the method* seaborn.kdeplot()*with the help of kernel density estimation(kde) diagram. Here we are comparing the evaluation with respect to turnover and non-turnover. And gave a title ‘*Employee Evaluation Distribution – Turnover V.S. No Turnover*’ by using* matplotlib.pyplot() *method.*

*And the output of* In[31] *is,*

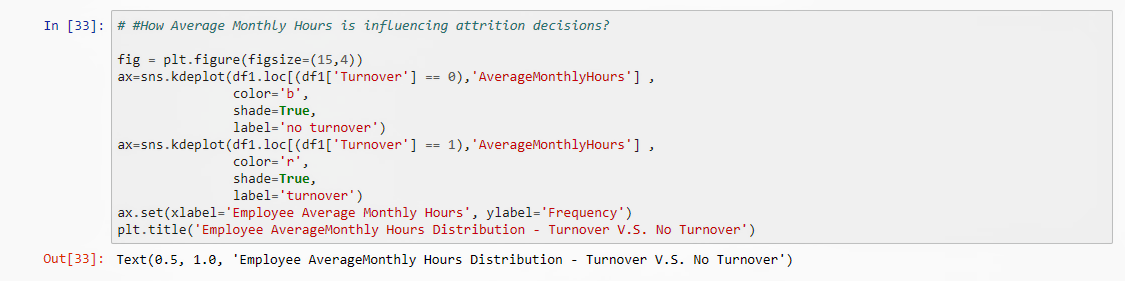
**

**

**

**

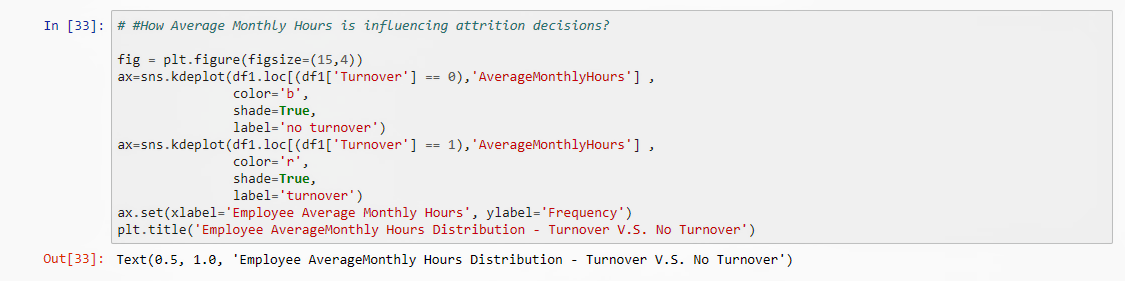
*Now in this line we have compared the mean employee turnover evaluation with the evaluation of the employee stayed in the company.*

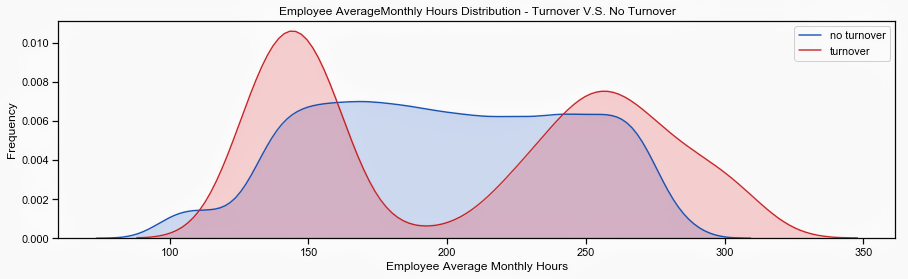
**

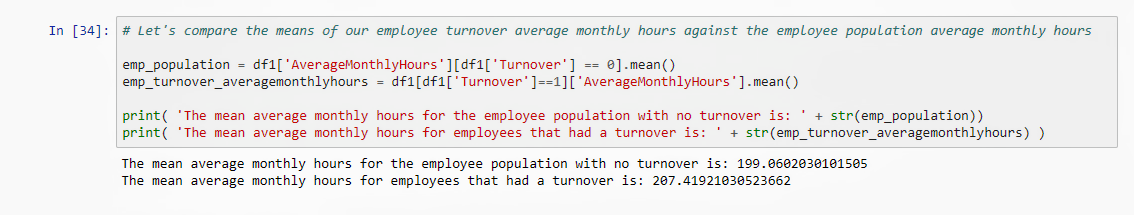
**

*Now we are visualizing how average monthly hours is influencing attrition decisions by using the method* seaborn.kdeplot()*with the help of kernel density estimation(kde) diagram. Here we are comparing the average monthly hours with respect to turnover and non-turnover. And gave a title ‘*Employee Average Monthly Hours Distribution – Turnover V.S. No Turnover*’ by using* matplotlib.pyplot() *method.*

*And the output of* In[33] *is,*

**

**

**

*Now in this line we have compared the mean employee turnover average monthly hours with the evaluation of the employee stayed in the company.*