Well, here is a short milestone report for my capstone project "IBM HR Analytics Employee Attrition & Performance" which includes:

**1)Define the problem:-**

As all of us know that Attrition in human resources refers to the gradual loss of employees over time. Job postings, hiring processes, paperwork and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

In general, relatively high attrition is problematic for companies. A major problem in high employee attrition is its cost to an organization. Employee attrition is not caused by one single factor for every worker in the organization but also involves multiple factors which need to be defined and analyzed further to determine the impact which leads to the high attrition rate.

**2)Identify your client:-**

Since we are dealing with IBM Employee Attrition dataset so our client is IBM here and they want to know the factors that lead to employee attrition. The client would like to know the contribution of varied factors and analysis to determine which factors contribute the most to employee attrition.

**3)Describe your data set, and how you cleaned/wrangled it:-**

My dataset consists of 35 columns/features which includes: Age, Attrition, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, Gender,  HourlyRate, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, MonthlyRate, NumCompaniesWorked, Over18, OverTime, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager and consist of ( int and object ) data type. An object is often an alias for a string.

**Steps involved in Cleaning/Wrangling data using Pandas and Jupyter notebook:**

***a)Importing and reading CSV file:*** I started by importing and reading CSV file with the help of pd.read\_csv() function.

***b)Finding missing values:*** While loading my dataset, I didn't find any missing value and my dataset is precleaned. But still, to double check that my dataset contains any missing value, I verified by using. isnull() function and chained it with .sum() function to get the count of any missing value in my data frame.

***c)Filling missing values:*** Since I didn't encounter any missing value, there was no such scenario of filling up the missing values.

***d)Checking the data types/shaping/describing:***

I also have to consider what type of values each of our columns are stored and summary of the basic information about the DataFrame and its data, to check this I used the following command: df1.info().

Next, I wanted to know how many columns and rows are present in our dataset. To do that I used .shape like df1.shape

I wanted to see some key stats in my data frame without delving too deep, mean values, min, and max so that we can get a feel for what we're working with. To do that I used the following command: df1.describe

Finally, after all these pieces of stuff, now my data was ready to deal with some sort of manipulation, filtering and merging/concatenating/joining depending upon the requirement.

**4)Explain your initial findings:**

My initial findings on my dataset were:

*1)Which age-group people contribute maximum attrition?*

I first started my analysis by surveying "Age" feature of my dataset.I explored ten most working-age people in the company by using value\_counts().I also plotted bar graph and found that people of Age 34 and 35 are topmost ages employed. Also, the mean age of the people is 37 which is quite close to the top age working employees in the company.This means most of the people belong to mid-age adults and it can be one of the factors for higher attrition.

So, I further deep dived into that and divided age into four groups:

(Age 15-24): Young\_adults

(Age 25-40): Mid\_age\_adults

(Age 41-54): Mid\_to\_old\_adults

(Age 55-64): Old\_adults

Also counted the number of Attrition('No'/'Yes') for each group and plotted bar graph. From the graph, I found that count of Mid\_age\_adults are more and are the highest contributor to Attrition and Old\_adults group of people have least attrition rate.

*2)What is the count of married people and unmarried people attrition rate? Are married people more prone to attrition?*

I found that based on MaritalStatus, people who are 'Single' are more prone to attrition than people who are married or divorced which is another interesting fact.

*3)What is the count of people working OverTime and YearsInCurrentRole? How working overtime (or not), and the years in role relate to employee attrition?*

First I counted the number of employees who have spent the same number of years in their current role.I analyzed that maximum number of employees leaving the company has spent less than a year, between 2-3yrs or 7-8yrs in the company. I further included overtime feature in the above analysis to see if people who are working overtime and spending the same number of years in the company are leaving the company more frequently or not.After plotting factor plot graph, I observed that people who worked overtime have left the company most.

4)What is the count of attrition of each department on the basis of RelationshipSatisfaction?Does satisfaction level has any impact on employees leaving these department?

First I counted attrition on each department by using count plot and observed that research and development field has highest attrition count. Then I added another feature 'RelationshipSatisfaction' to see if the satisfaction level has any impact on employees leaving these departments. After plotting factor plot, I don't see any variance on attrition based on the RelationshipSatisfaction feature. All four different levels(Low, Medium, High, Very High) has the similar number of attrition.

5)Do JobSatisfaction and JobRole impact gradual loss of employees? Are these two features have a common pattern?

From the plotted graph based on JobRole and JobSatisfaction feature, I see that employees in these roles(Sales Executive, Research Scientist, Laboratory Technician) leave the company most if they have low job satisfaction or high job satisfaction(which is not very high). Sales Representative employees count of attrition are very similar under all job satisfaction levels.

6)At last, I am wrapped up by plotting each variable and seeing its influence on the Attrition of the organization.