**A**

**PROJECT REPORT**

**On**

**DATA SCIENCE PROCESS TO SOLVE EMPLOYEE ATTRITION**



**Submitted by**

**Krishna chetri** (ET16BTHEC031)

**Pranjit das** (ET17BTHEE024L)

**Mandeep chetry** (ET16BTHC42)

**Pragyanav saikia** (ET16BTHEC033)

**Sahidul islam** (ET16BTHEC0)

**School of Engineering and Technology**

**Department Of Computer Science and Engineering**

**The Assam Kaziranga University**

# 

# 

# CONTENTS

# INTRODUCTION

## **FLOW DIAGRAM**

# INCLUDED COMPONENTS DETAILS

# STEPS

# SAMPLE OUTPUT OF THE PROGRAMMING CODE

# CONCLUSION

# INTRODUCTION

# This code pattern is a high-level overview of what to expect in a data science pipeline and the tools that can be used along the way. It starts from framing the business question, to building and deploying a data model. The pipeline is demonstrated through the employee attrition problem.

Employees are the backbone of any organization. Its performance is heavily based on the quality of the employees and retaining them. With employee attrition, organizations are faced with a number of challenges:

1. Expensive in terms of both money and time to train new employees
2. Loss of experienced employees
3. Impact on productivity
4. Impact on profit

The following solution is designed to help address the employee attrition problem. When the reader has completed this code pattern, they will understand:

1. The Process involved in solving a data science problem
2. How to create and use Watson Studio instance
3. How to mitigate bias by transforming the original dataset through use of the AI Fairness 360 (AIF360) toolkit
4. How to build and deploy the model in Watson Studio using various tools

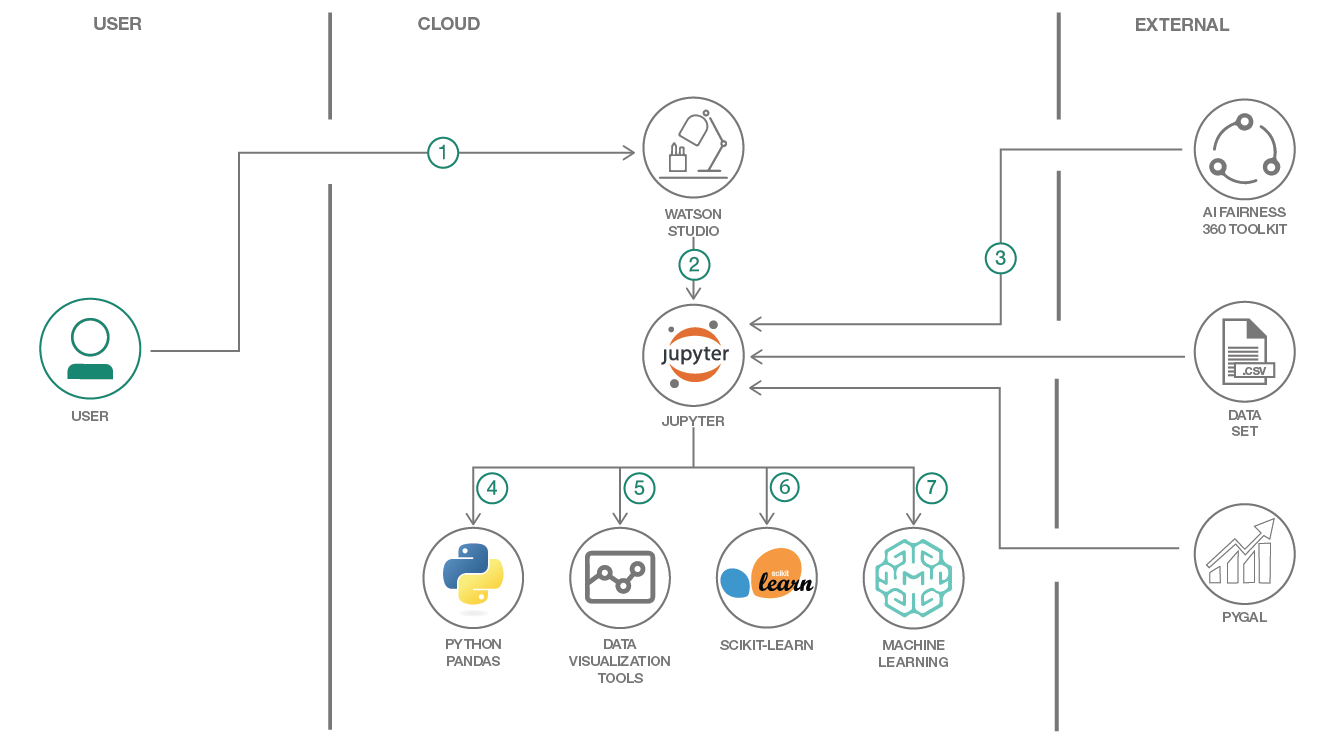
The dataset used in the code pattern is supplied by [Kaggle](https://www.kaggle.com/) and contains HR analytics data of employees that stay and leave. The types of data include metrics such as education level, job satisfactions, and commmute distance.

# The data is made available under the following license agreements:

### Dataset License Details

| **Asset** | **License** | **Source Link** |
| --- | --- | --- |
| [Employee Attrition Data - Database License](https://github.com/IBM/employee-attrition-aif360/blob/master/data/emp_attrition.csv) | [Open Database License (ODbL)](https://opendatacommons.org/licenses/odbl/1.0/) | [Kaggle](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/home) |
| [Employee Attrition Data - Content License](https://github.com/IBM/employee-attrition-aif360/blob/master/data/emp_attrition.csv) | [Database Content license (DbCL)](https://opendatacommons.org/licenses/dbcl/1.0/) | [Kaggle](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/home) |

## **FLOW DIAGRAM**



1. Create and login to the IBM Watson Studio.
2. Upload the jupyter notebook and start running it.
3. Notebook downloads the dataset and imports fairness toolkit (AIF360) and Pygal data visualization library.
4. Pandas is used for reading the data and perform initial data exploration.
5. Matplotlib, Seaborn, Plotly, Bokeh and Pygal (from step-3) are used for visualizing the data.
6. Scikit-Learn and AIF360 (from step-3) are used for model development.
7. Use the IBM Watson Machine Learning feature to deploy and access the model to generate employee attrition classification.

**INCLUDED COMPONENTS DETAILS**

* [IBM Watson Studio](https://www.ibm.com/bs-en/marketplace/data-science-experience): Analyze data using RStudio, Jupyter, and Python in a configured, collaborative environment that includes IBM value-adds, such as managed Spark.
* [IBM Watson Machine Learning](https://cloud.ibm.com/catalog/services/machine-learning): a set of REST APIs to develop applications that make smarter decisions, solve tough problems, and improve user outcomes.
* [Jupyter Notebook](http://jupyter.org/): An open source web application that allows you to create and share documents that contain live code, equations, visualizations, and explanatory text.

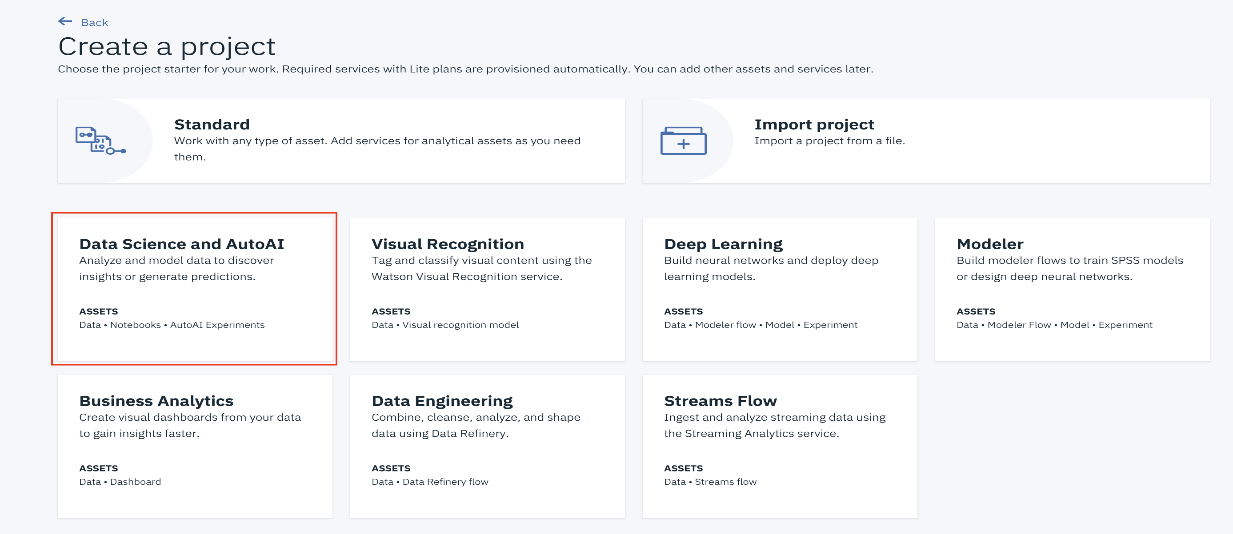
## **Featured technologies:**

* [Artificial Intelligence](https://medium.com/ibm-watson): Artificial intelligence can be applied to disparate solution spaces to deliver disruptive technologies.
* [Data Science](https://medium.com/ibm-watson): Systems and scientific methods to analyze structured and unstructured data in order to extract knowledge and insights.
* [Python](https://www.python.org/): Python is a programming language that lets you work more quickly and integrate your systems more effectively.
* [Pandas](http://pandas.pydata.org/): A Python library providing high-performance, easy-to-use data structures.
* [AIF360 Fairness toolkit](http://aif360.mybluemix.net/): This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle.
* [Scikit-Learn](https://scikit-learn.org/stable/): Free software machine learning library for the Python programming language.
* Data Visualization tools: Bokeh, Matplotlib, Seaborn, Pygal and Plotly.

## **STEPS**

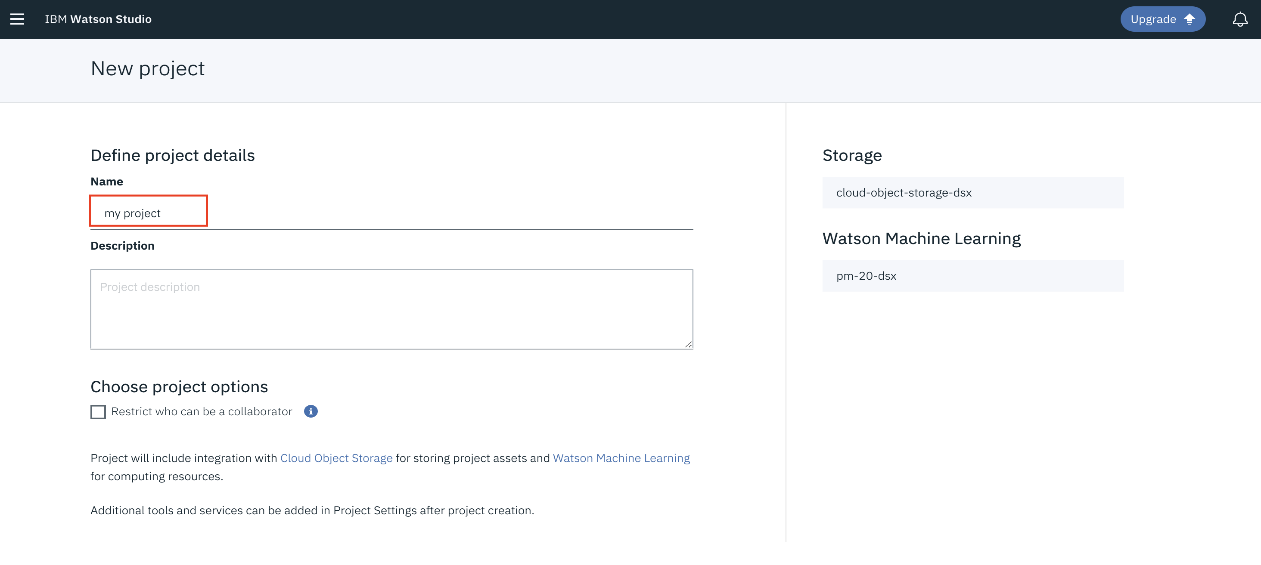
1. [Create a new Watson Studio project](https://github.com/IBM/employee-attrition-aif360#1-create-a-new-watson-studio-project)
2. [Create the notebook](https://github.com/IBM/employee-attrition-aif360#2-create-the-notebook)
3. Add Data Files
4. [Run the notebook](https://github.com/IBM/employee-attrition-aif360#3-run-the-notebook)
   * 1. **Create a new Watson Studio project**:

* Log into IBM's [Watson Studio](https://dataplatform.cloud.ibm.com/). Once in, you'll land on the dashboard.
* Create a new project by clicking + New project and choosing Data Science:

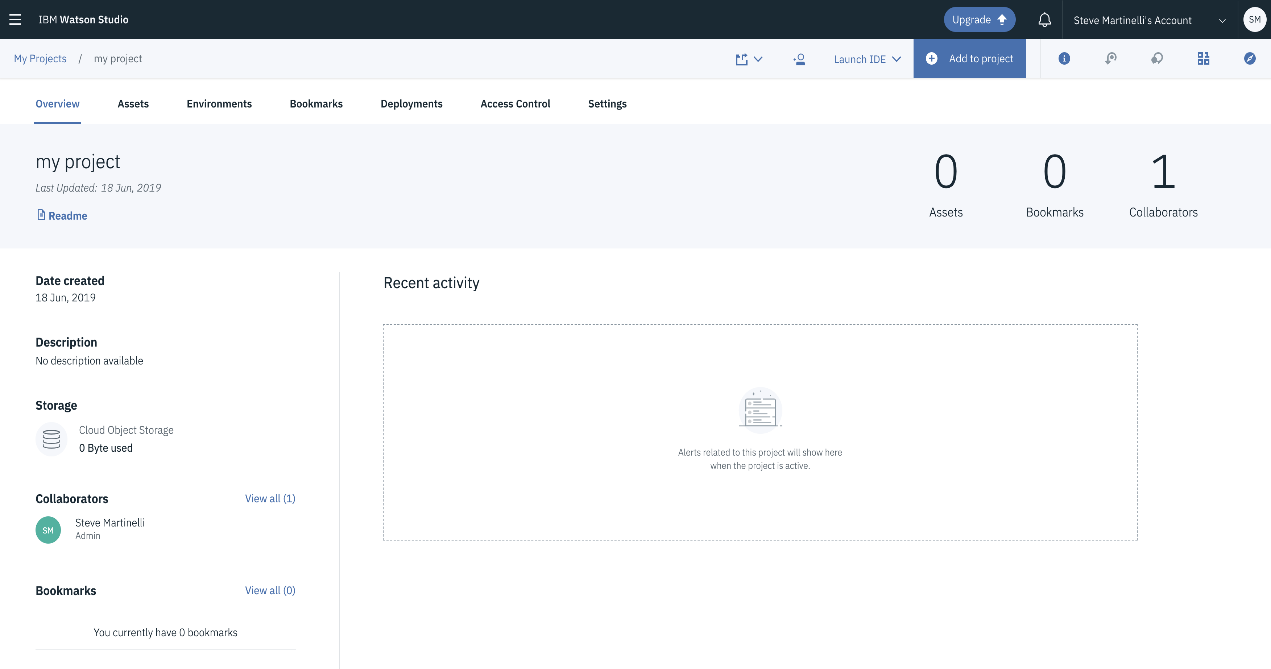


* Enter a name for the project name and click Create:

**NOTE**: By creating a project in Watson Studio a free tier Object Storage service and Watson Machine Learning service will be created in your IBM Cloud account. Select the Free storage type to avoid fees.



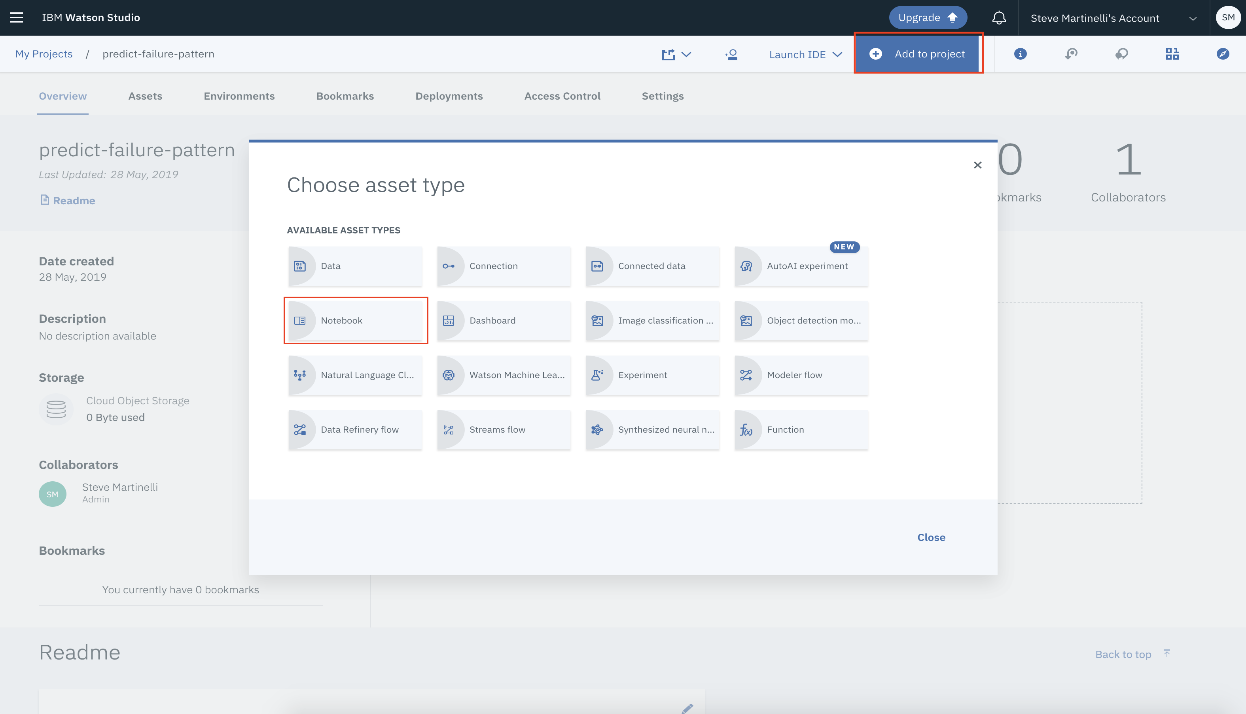
* Upon a successful project creation, you are taken to a dashboard view of your project. Take note of the Assets and Settings tabs, we'll be using them to associate our project with any external assets (datasets and notebooks) and any IBM cloud services.



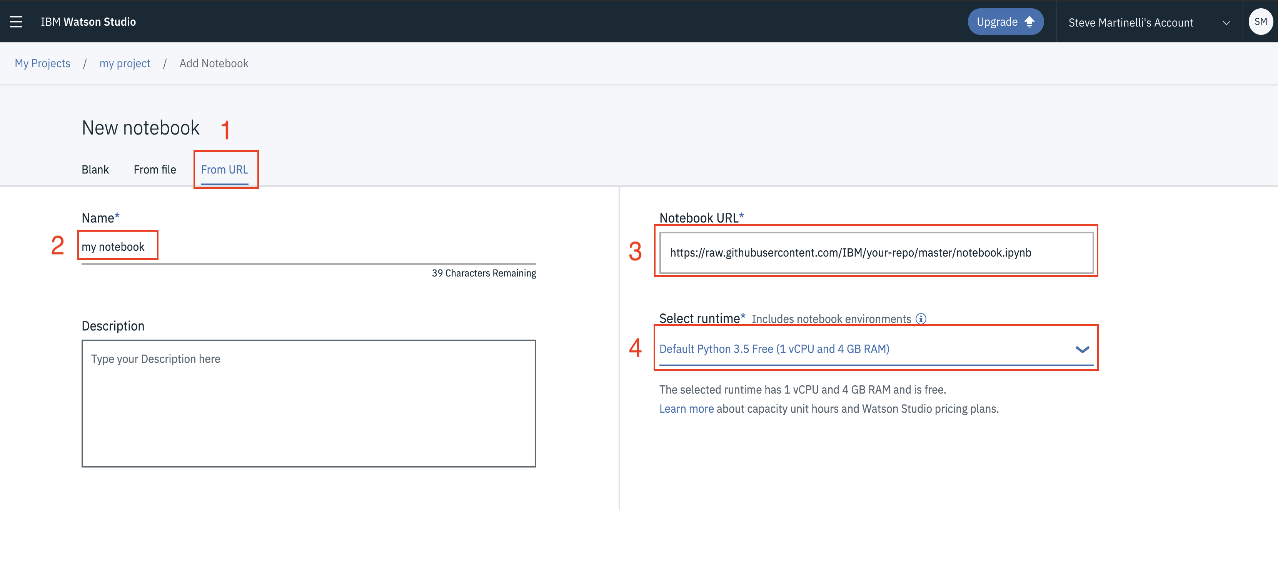
## **Create the Notebook:**

The notebook we'll be using can be viewed in [notebooks/employee-attrition.ipynb](https://github.com/IBM/employee-attrition-aif360/blob/master/notebooks/employee-attrition.ipynb), and a completed version can be found in [examples/employee-attrition.ipynb](https://github.com/IBM/employee-attrition-aif360/blob/master/examples/employee-attrition.ipynb).

* From the new project Overview panel, click + Add to project on the top right and choose the Notebook asset type.



* Fill in the following information:
* Select the From URL tab. [1]
* Enter a Name for the notebook and optionally a description. [2]
* Under Notebook URL provide the following url: <https://github.com/IBM/employee-attrition-aif360/blob/master/notebooks/employee-attrition.ipynb> [3]
* For Runtime select the Python 3.5 option. [4]

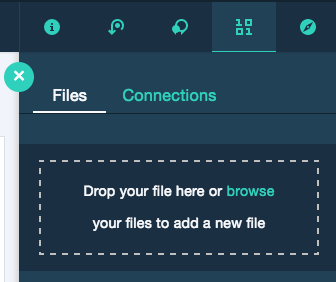


**TIP:** Once successfully imported, the notebook should appear in the Notebooks section of the Assets tab.

### **Add data files:**

## Add **README.md** and **Pipeline\_LabelEncoder-0.1.zip** file, created/downloaded in [this section](https://github.com/IBM/image-classification-using-cnn-and-keras#21-data-preparation), to Object Storage. In Watson Studio, go to your project default page, use Find and Add Data (look for the 10/01 icon) and its Files tab

* Click browse and upload **README.md** and **Pipeline\_LabelEncoder-0.1.zip** file.



If you use your own dataset, you will need to update the variables/folder names that refer to the data files in the Jupyter Notebook.

To open the notebook, click on the edit icon to start editing the notebook on your project.

* 1. **Run The Notebook:**

When running the notebook, you will come to the cell that requires you to enter your Watson Machine Learning instance credentials. These will be required to complete the notebook. Refer to step #1 above for more details.

When a notebook is executed, what is actually happening is that each code cell in the notebook is executed, in order, from top to bottom.

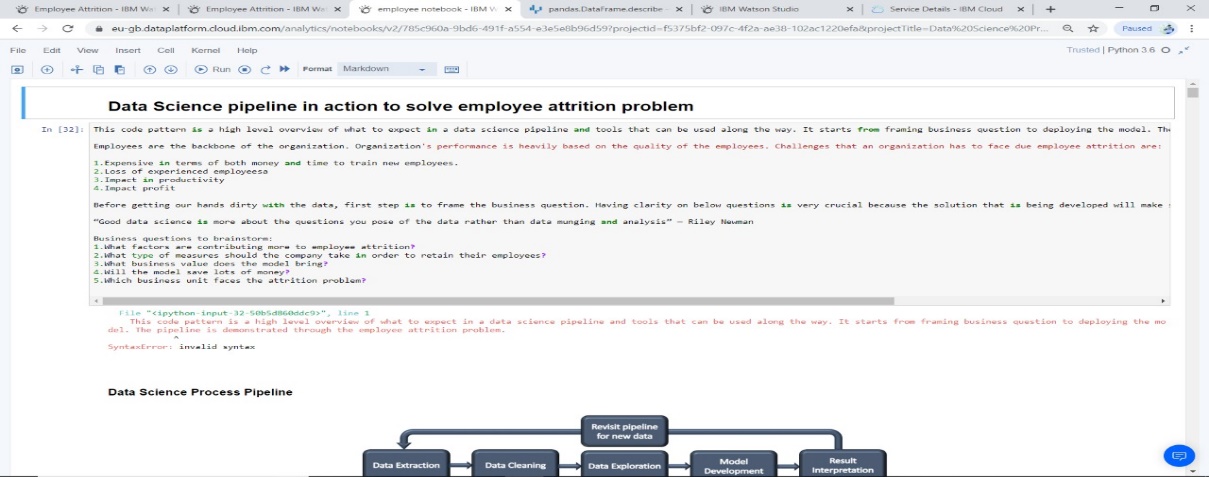
Each code cell is selectable and is preceded by a tag in the left margin. The tag format is In [x]: Depending on the state of the notebook, the x can be:

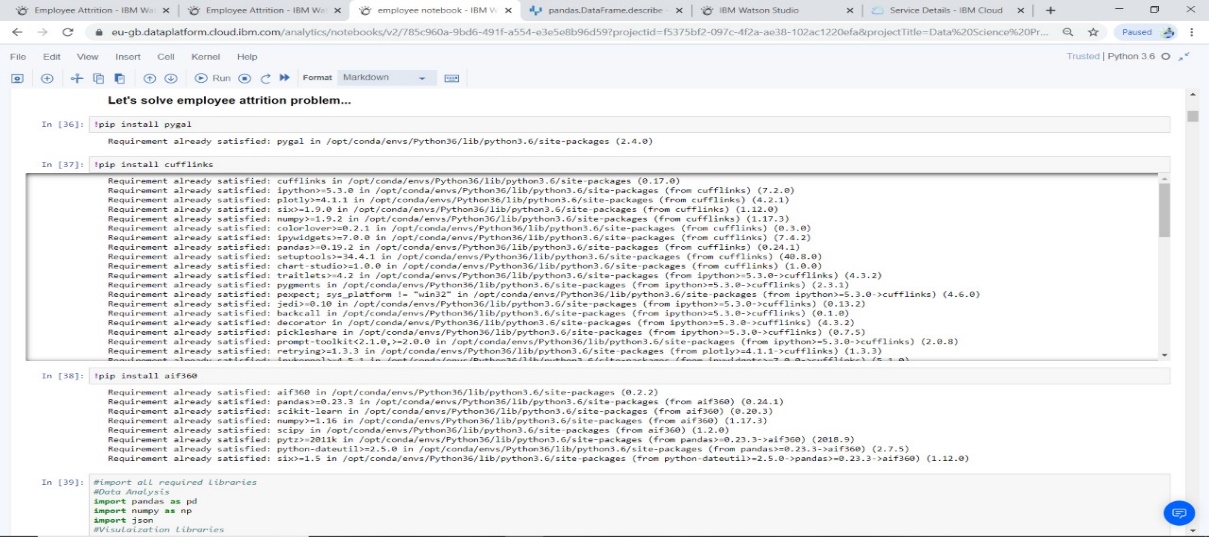
* A blank, this indicates that the cell has never been executed.
* A number, this number represents the relative order this code step was executed.
* A \*, this indicates that the cell is currently executing.

There are several ways to execute the code cells in your notebook:

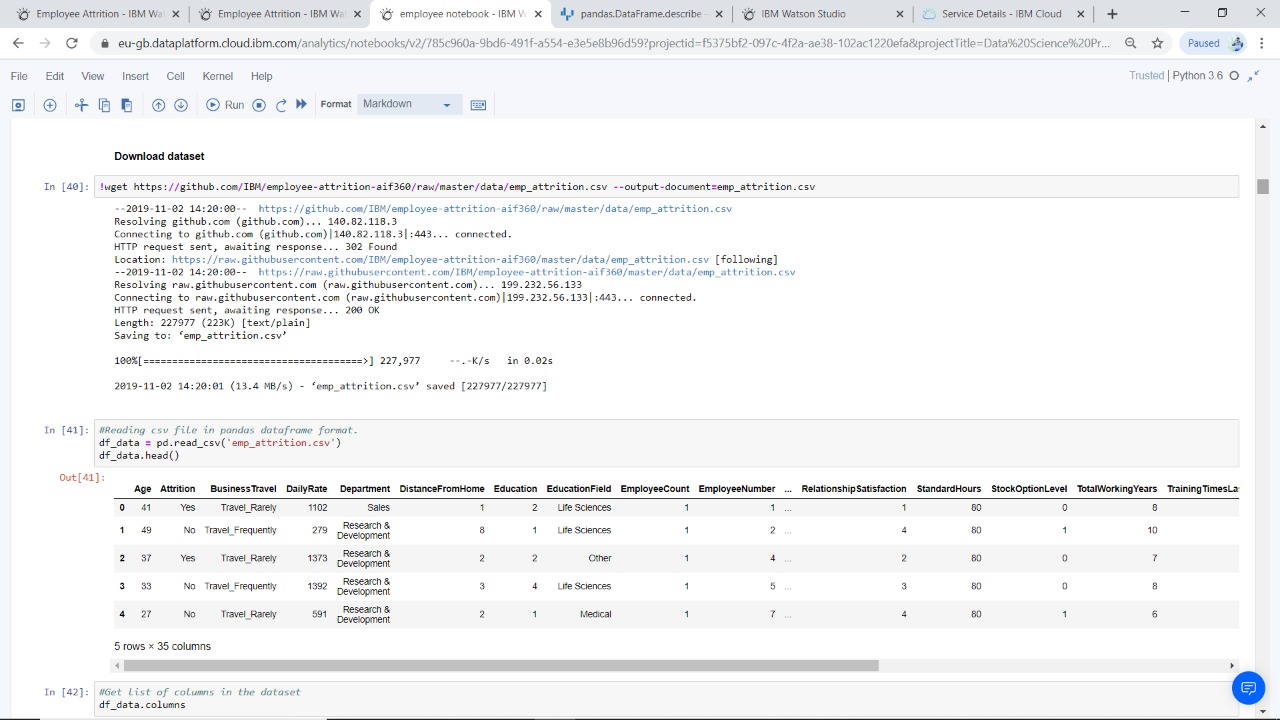
* One cell at a time.
  + Select the cell, and then press the Play button in the toolbar.
* Batch mode, in sequential order.
  + From the Cell menu bar, there are several options available. For example, you can Run All cells in your notebook, or you can Run All Below, that will start executing from the first cell under the currently selected cell, and then continue executing all cells that follow.
* At a scheduled time.
  + Press the Schedule button located in the top right section of your notebook panel. Here you can schedule your notebook to be executed once at some future time, or repeatedly at your specified interval.

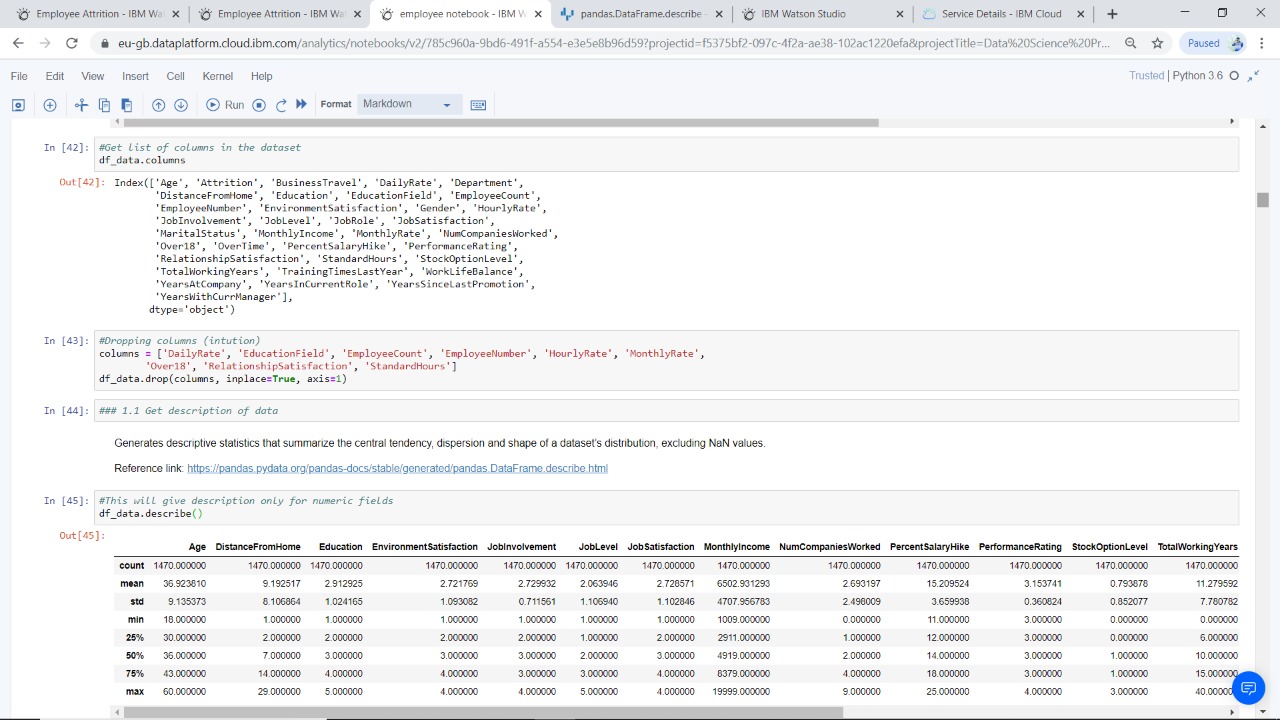
## **Sample output of the Programming code**

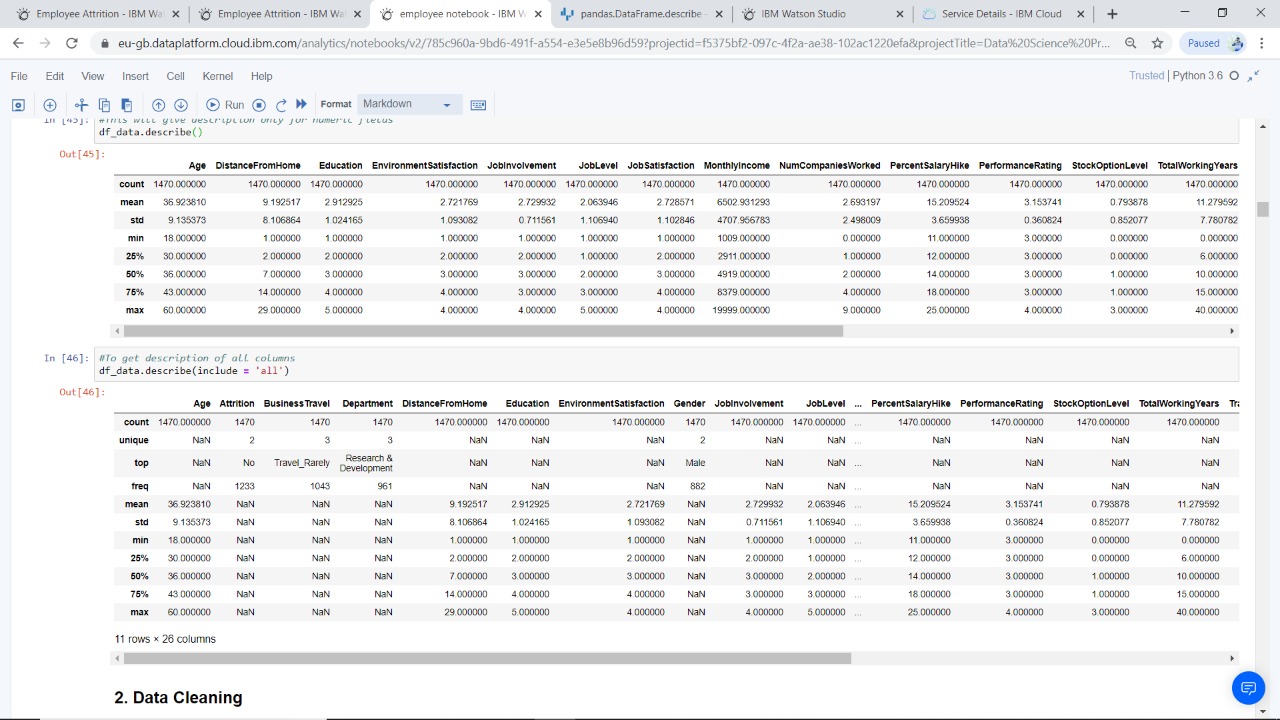
1. 

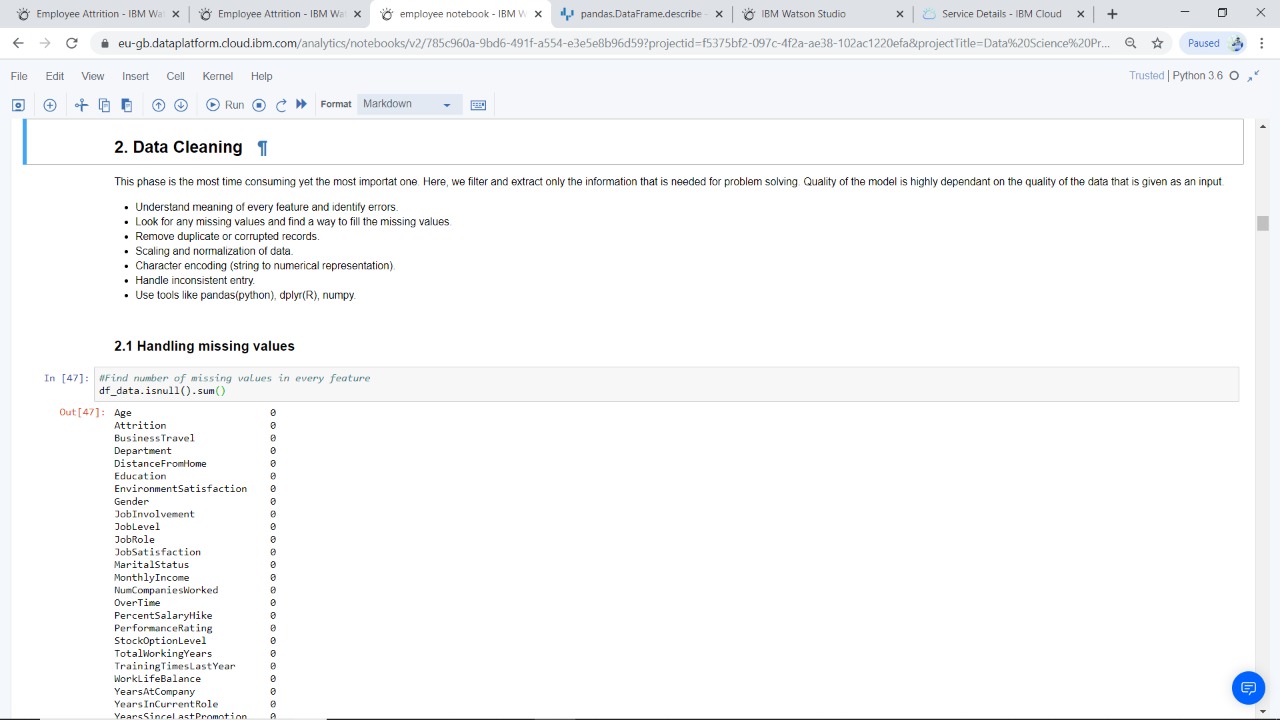
2. 

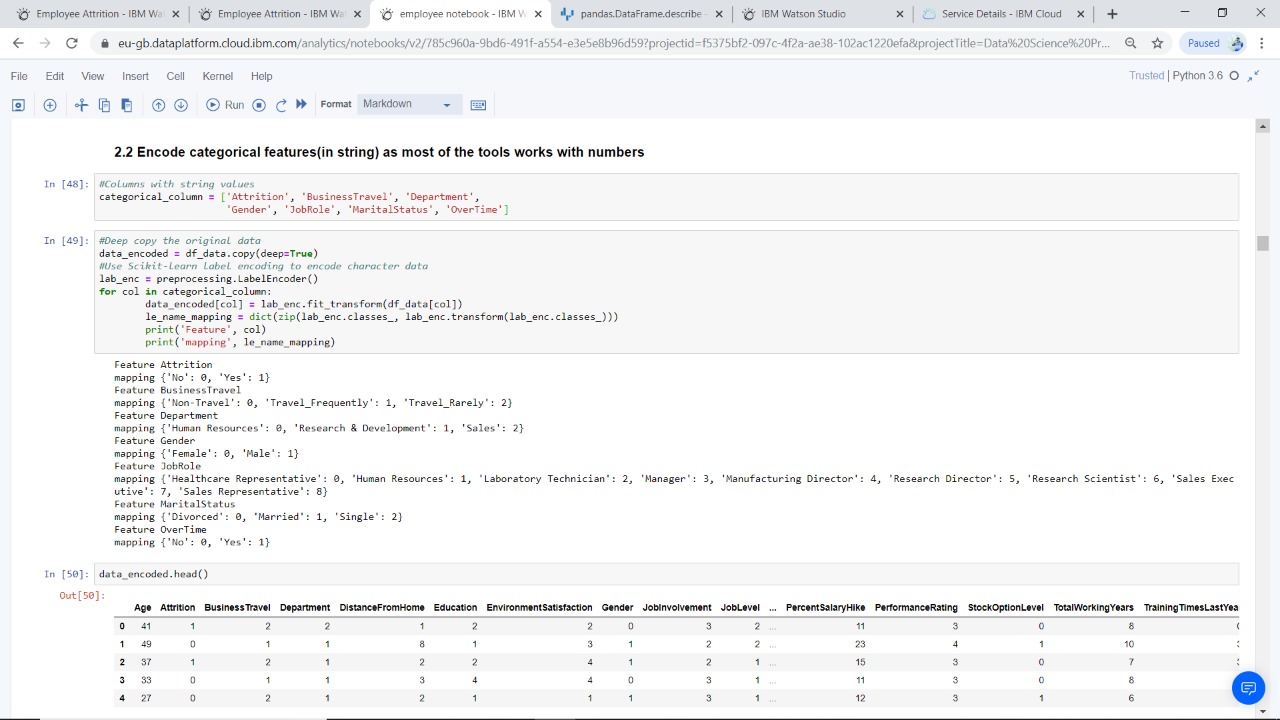
## 3. C:\Users\LENOVO\Desktop\ibm photos\WhatsApp Image 2019-11-10 at 5.27.30 PM.jpeg

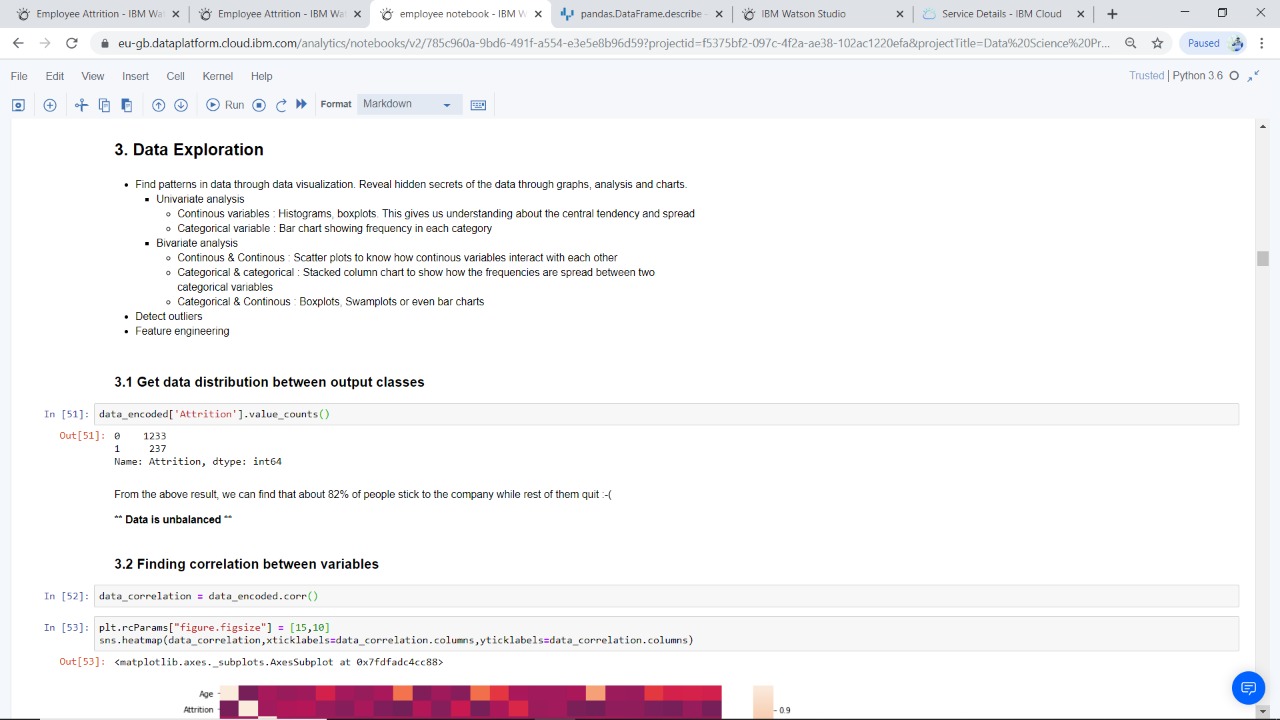
4.

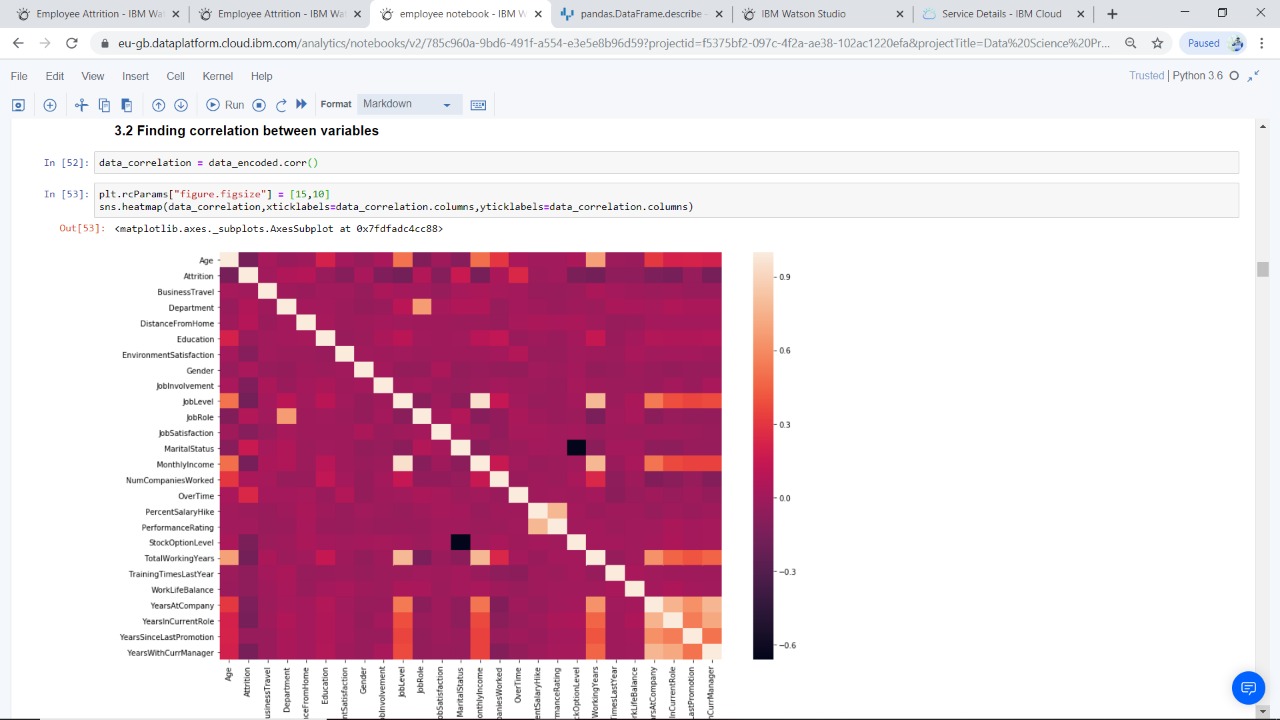
5.

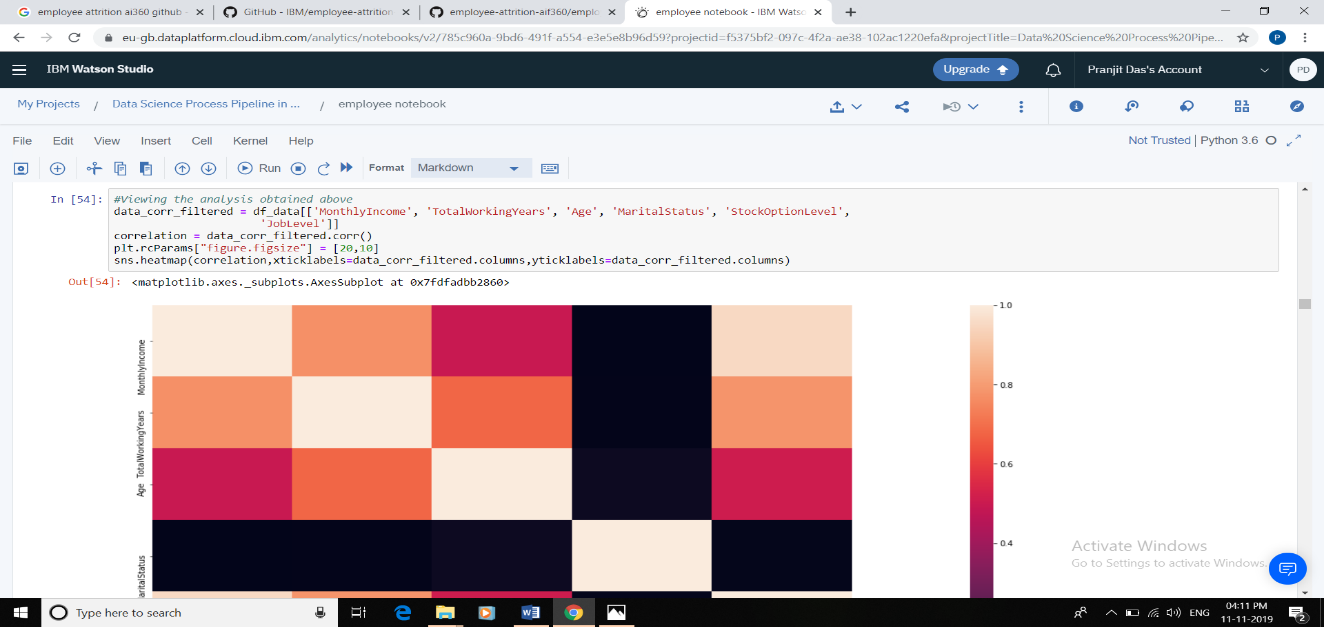
6.

7.

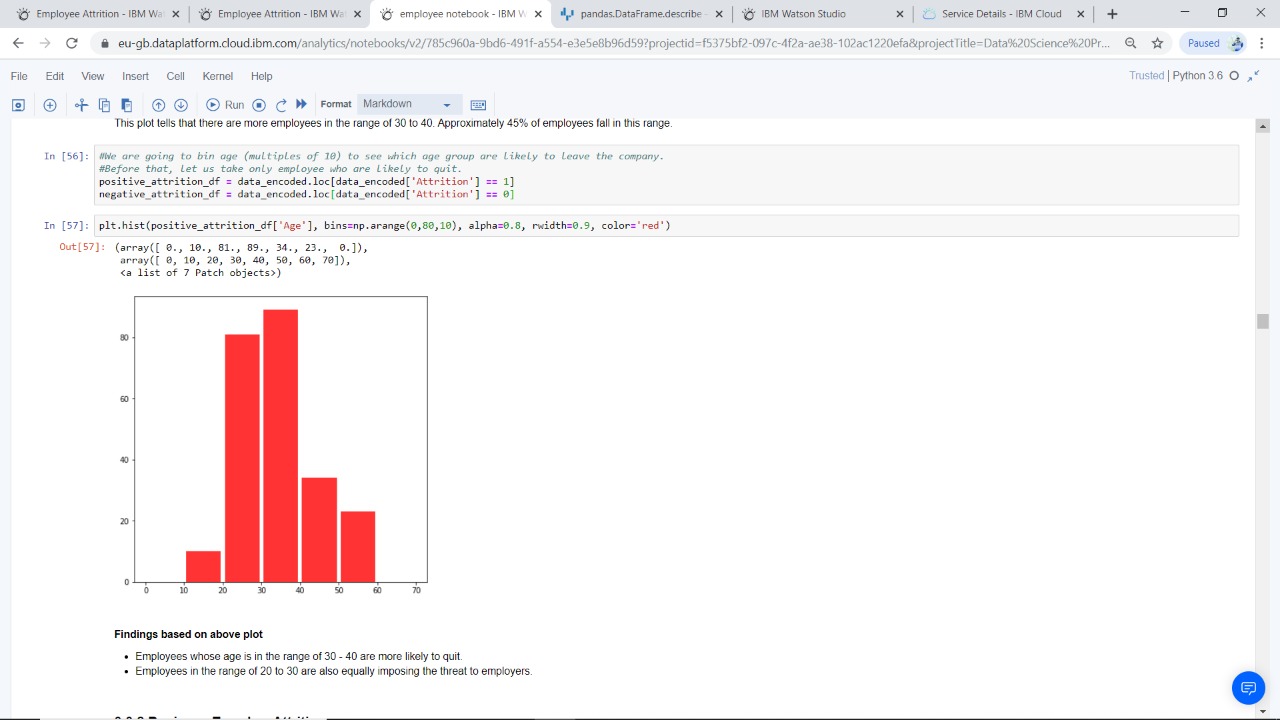
8.

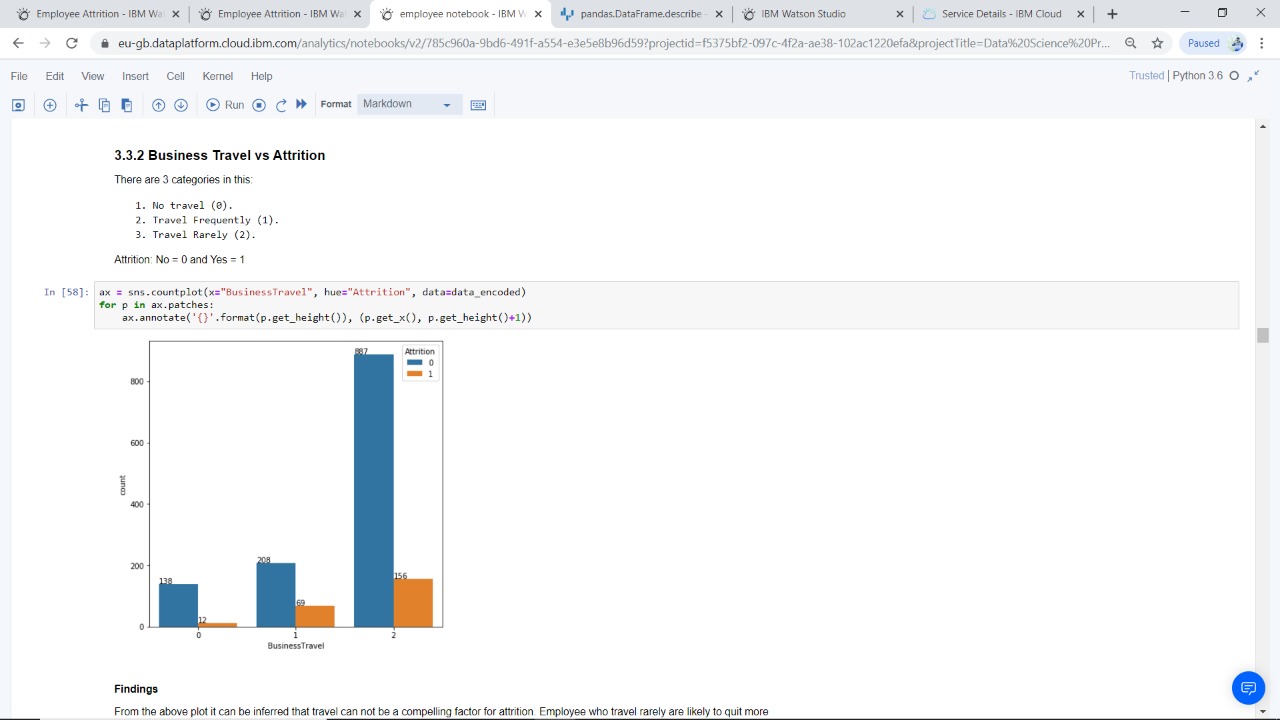
9.

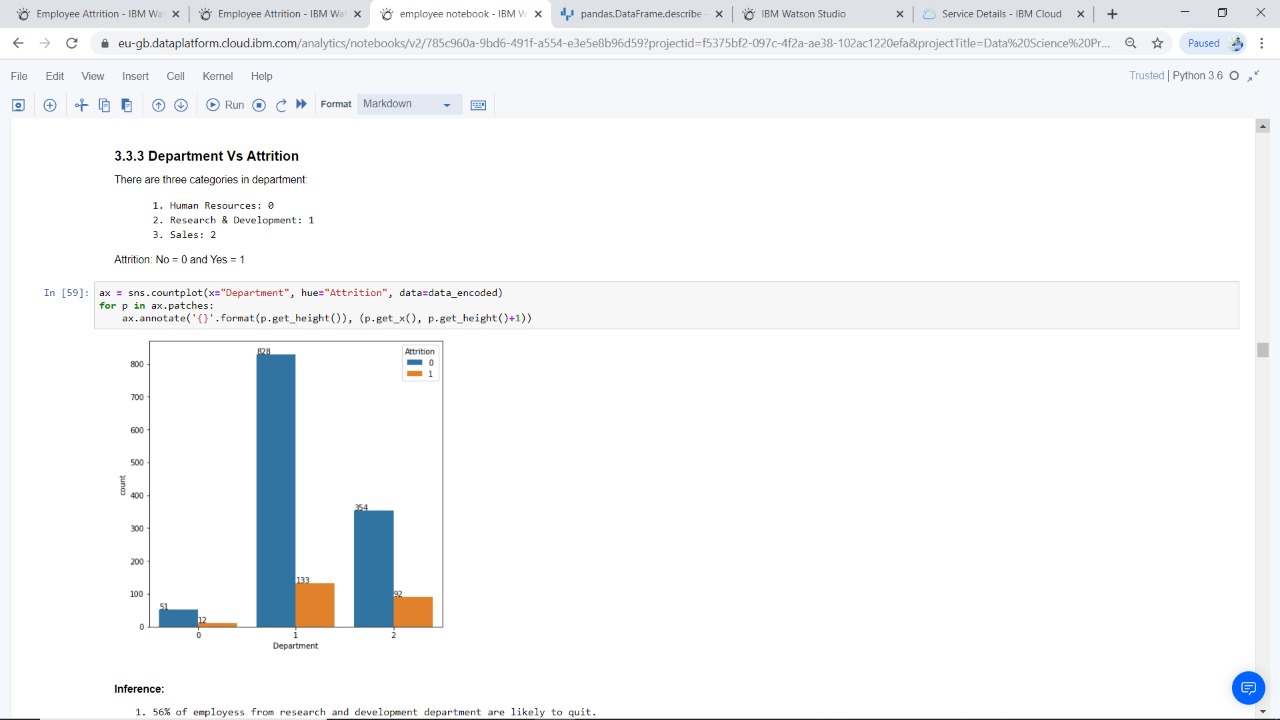
10.

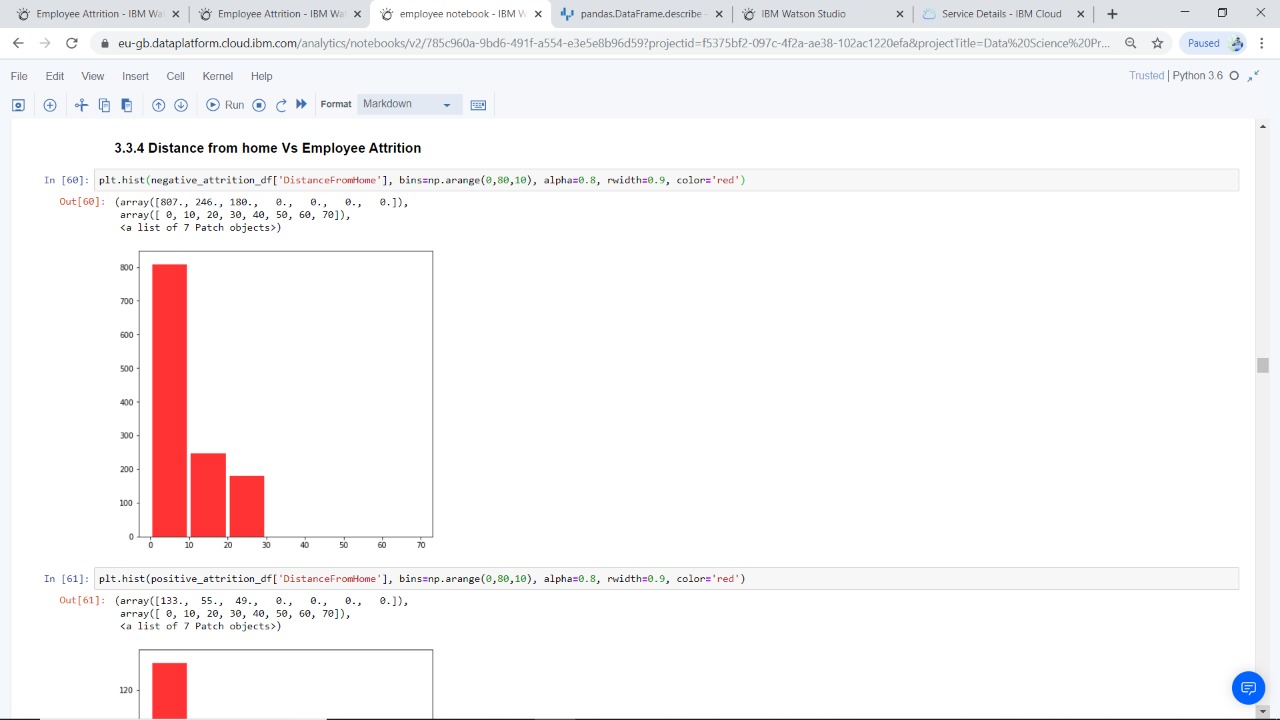
11.

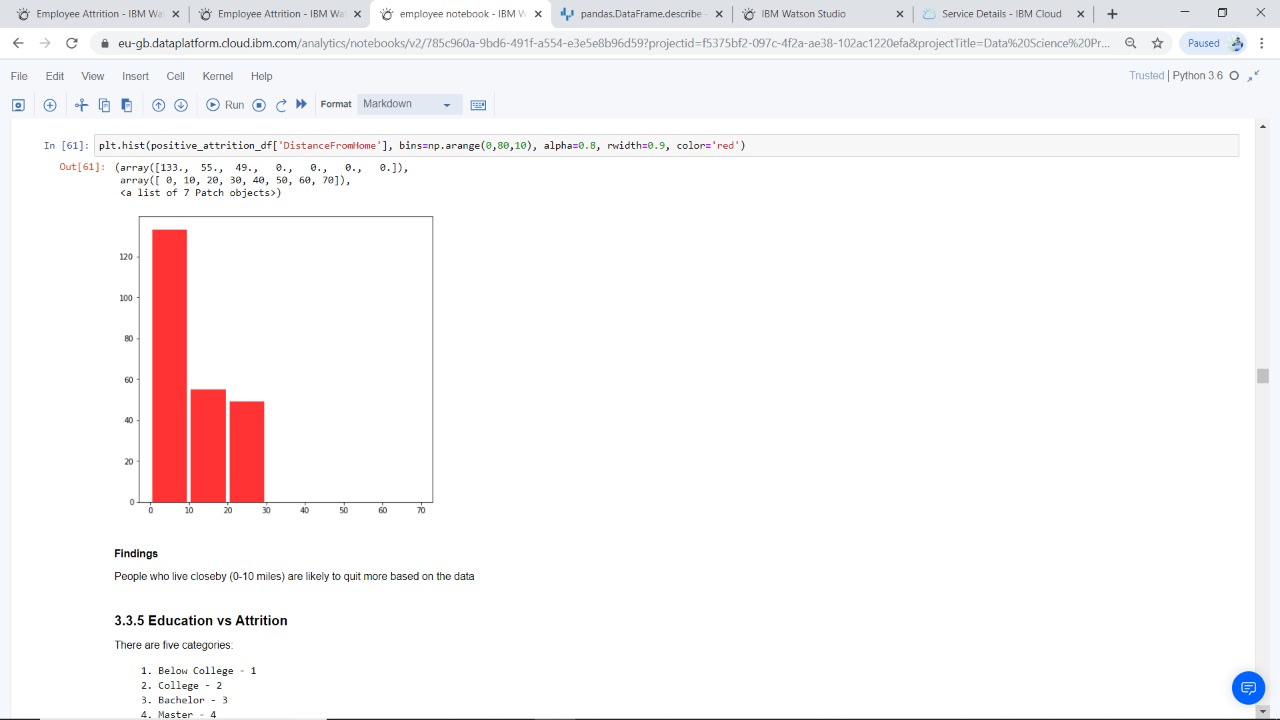
12.

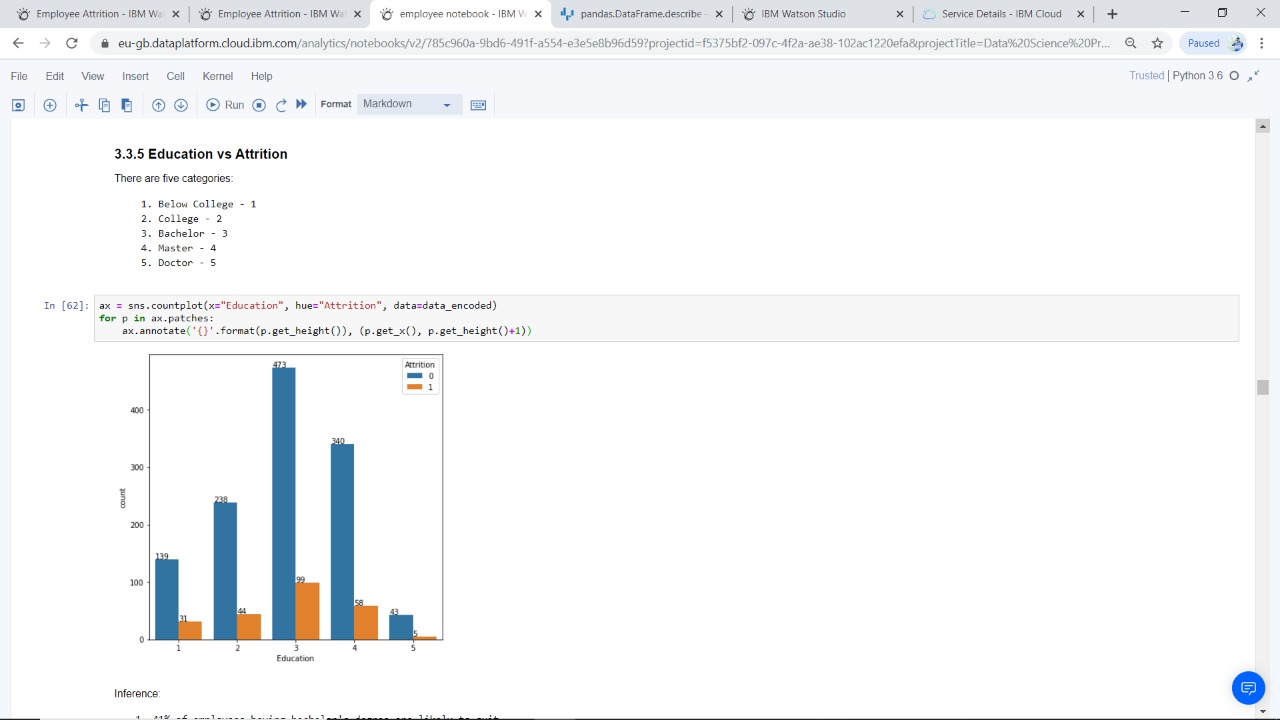
13.

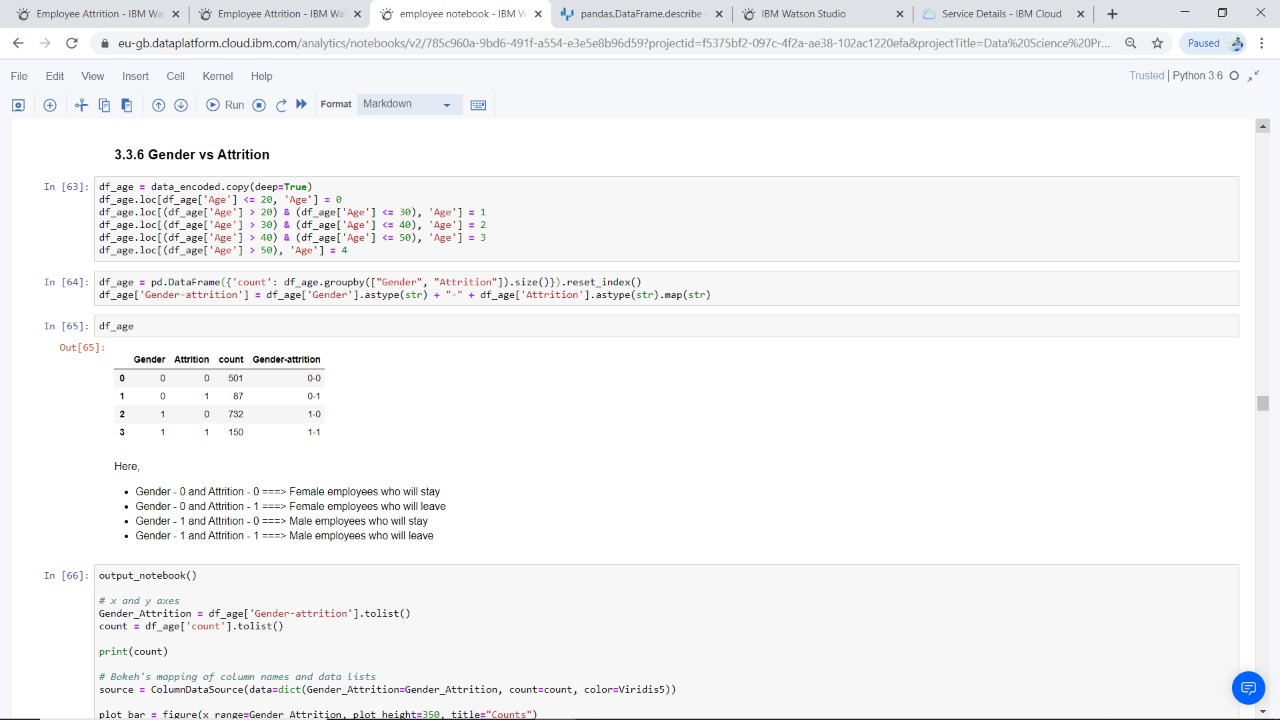
14.

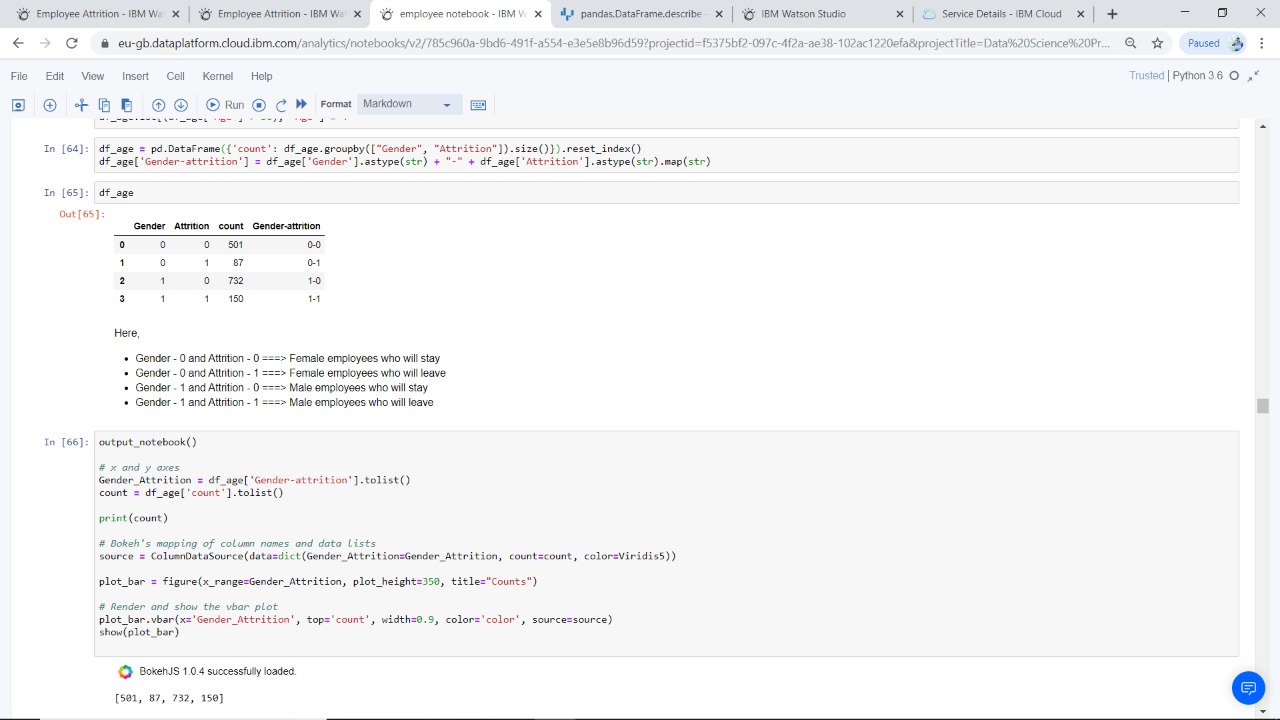
15.

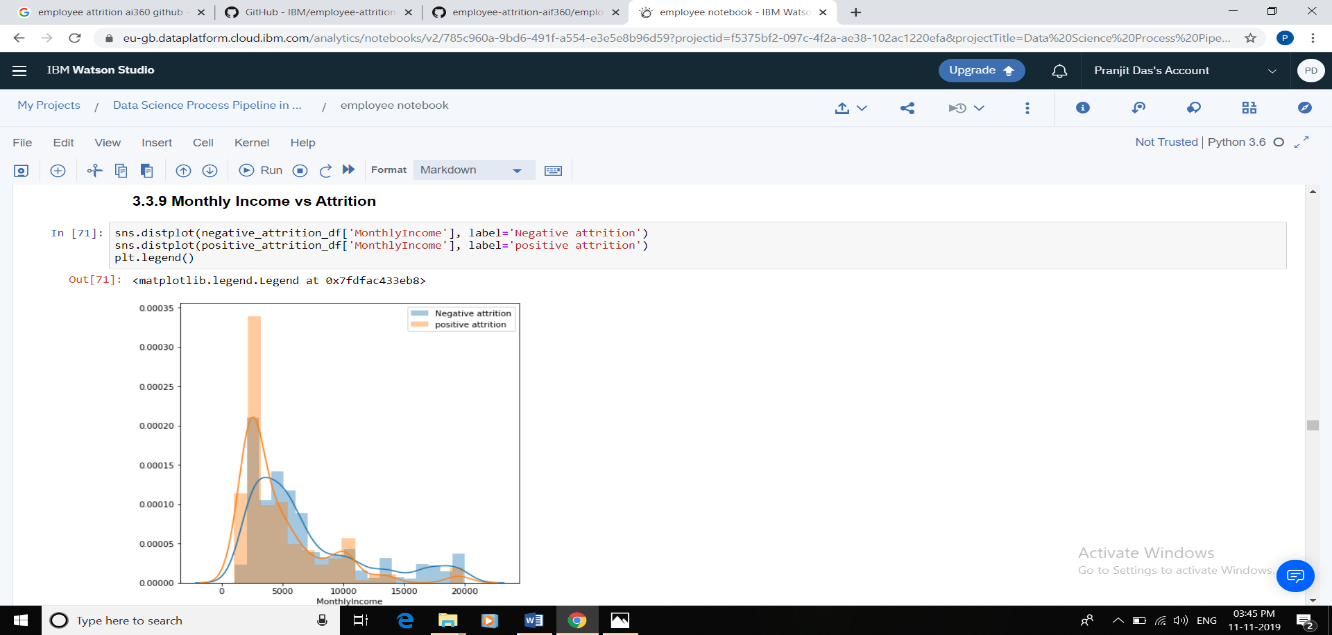
16.

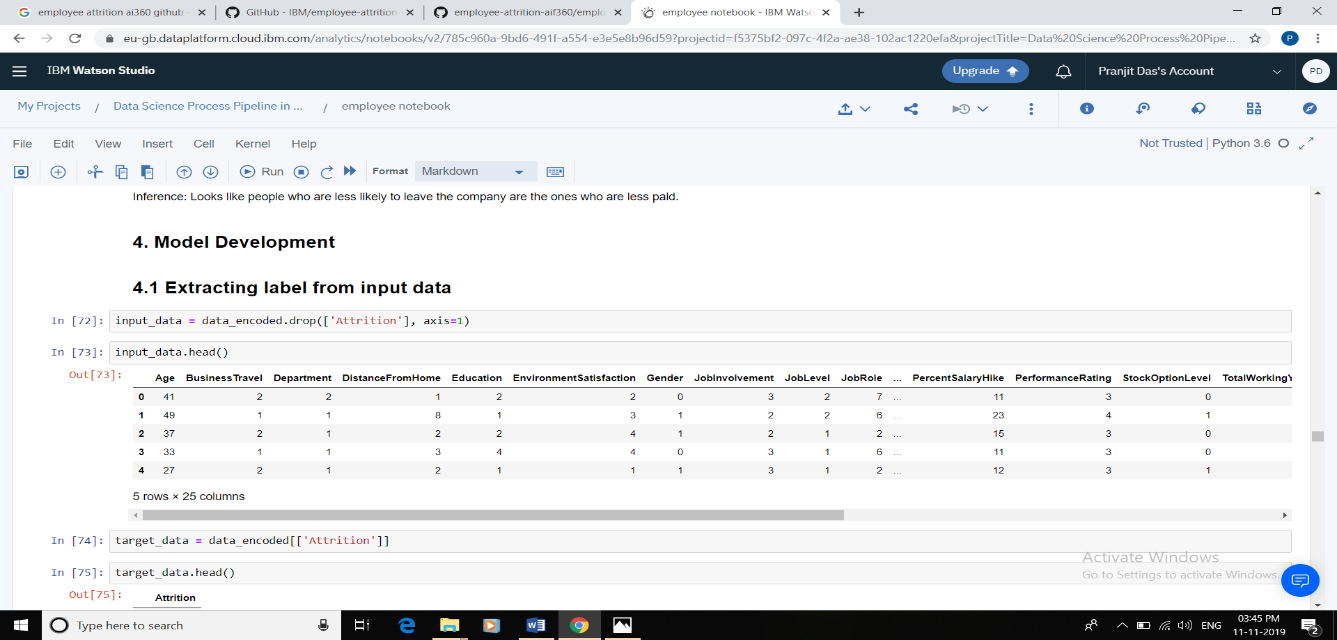
17.

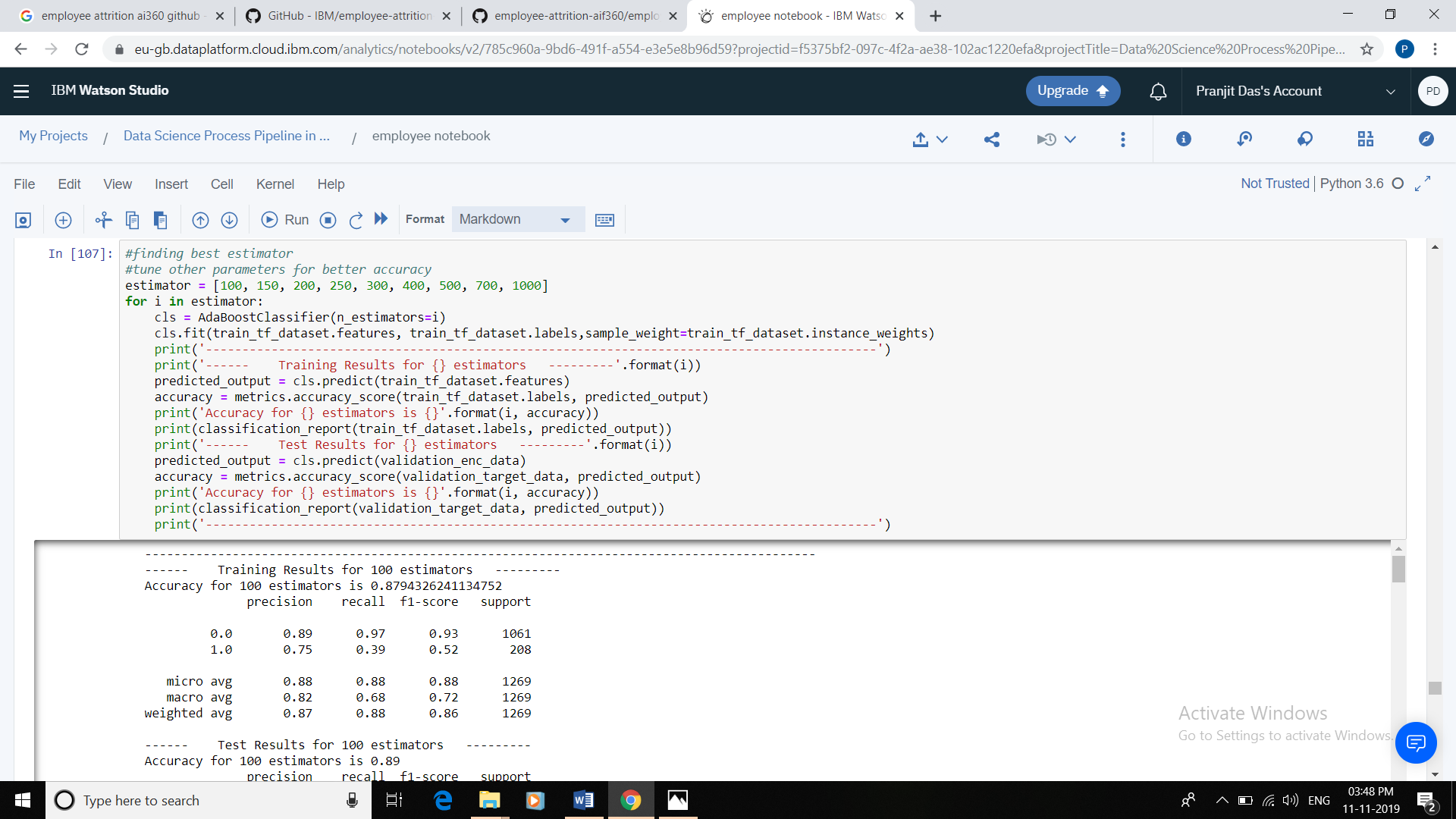
18.

19.

20.

21.

22.

23.

**CONCLUSION**

1. Factors contributing more to the employee attrition are Monthly Income, Total Working Years, Years At Company, Years In Current Role, Years With Current Manager, Age, Over Time, Distance From Home, Stock Option Level, Job Level, Job Role, Work Life Balance, Gender.
2. Top three roles facing attrition
   * 26% of employees who are likely to quit belong to Laboratory Technician group.
   * 24% of employees belong to Sales Executive group.
   * 19% of employees belong to Research Scientist group. (other inferences are mentioned below each graph)
3. The model developed will be able to predict whether an employee will stay or not. This will help company to know the status of an  
   employee in advance and take necessary actions to prevent loss that will incur.