**Student Name: Student Id:**

**Lab Instructions:**Please show your work to the instructor present in the lab before submitting.

**Submission Due**: End of laboratory class, submit the file on Moodle at least 10 minutes before the end of laboratory class.

**Total Marks** = 10 marks

Marks will be given only to students who attend and participate during 2-hour laboratory class. Submission on Moodle is mandatory as an evidence of participation.

**Learning Outcomes:**

|  |  |
| --- | --- |
| LO4 | Plan and Analyze Big Data |

**Marking Criteria:**

|  |  |  |
| --- | --- | --- |
| **Task Details** | **Submission Requirements** | **Marks** |
| **Lab Task 1-10:** Complete the following tasks using Pandas library in Python. | Same **Word file** containing the transcript of the answer.  Also submit **Python Notebook** file along with. | 10 (1 mark each) |

**Lab 4 – Data Analysis with Pandas Library**

**Introduction:**

[**Pandas**](http://pandas.pydata.org/) is a Python library that provides extensive means for data analysis. Data scientists often work with data stored in table formats like .csv, .tsv, or .xlsx. Pandas makes it very convenient to load, process, and analyze such tabular data using SQL-like queries. In conjunction with Matplotlib and Seaborn, Pandas provides a wide range of opportunities for visual analysis of tabular data.

The main data structures in Pandas are implemented with **Series** and **DataFrame** classes. The former is a one-dimensional indexed array of some fixed data type. The latter is a two-dimensional data structure - a table - where each column contains data of the same type. You can see it as a dictionary of Series instances. DataFrames are great for representing real data: rows correspond to instances (examples, observations, etc.), and columns correspond to features of these instances.

**Dataset:**

In this lab, we will understand how Pandas can be used to analyze big data in Python. The dataset used for this purpose is “[Churn in telecom’s dataset](https://bigml.com/user/francisco/gallery/dataset/5163ad540c0b5e5b22000383)” which is a telecommunication clients dataset available open source from UCI – Machine Learning Repository.

A “**churn**” in any business can be defined as a client who discontinues business with the organization. For example, a **telecom churn** would be a telecom line subscriber who discontinues using the service.

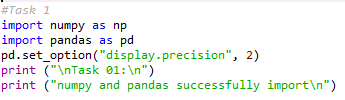
**Lab Exercise:**

1. Run **Jupyter Notebook** in your PC.
2. Open the notebook **“BI&BD\_SEN4210\_Lab4Ex\_05\_04\_2020”** in your Jupyter Notebook. The notebook is provided with lab content on Moodle.
3. Follow the instruction and perform the tasks to understand how you can do data analysis with Pandas.
4. The required dataset for this lab is available on Moodle

**Lab Tasks:** Complete the following tasks using Pandas library in Python.

1. Write code to import numpy and pandas.

**Code:**

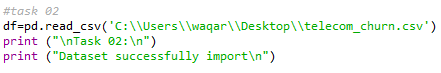


**Output:**



1. Write code to import the dataset as a pandas dataframe.

**Code:**

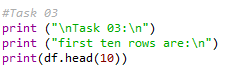


**Output:**

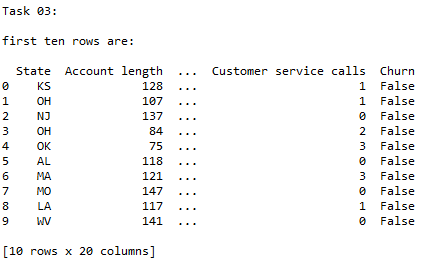


1. Write code to display first 10 rows of dataframe as head.

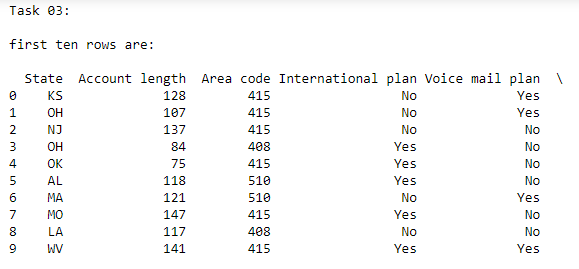
**Code:**

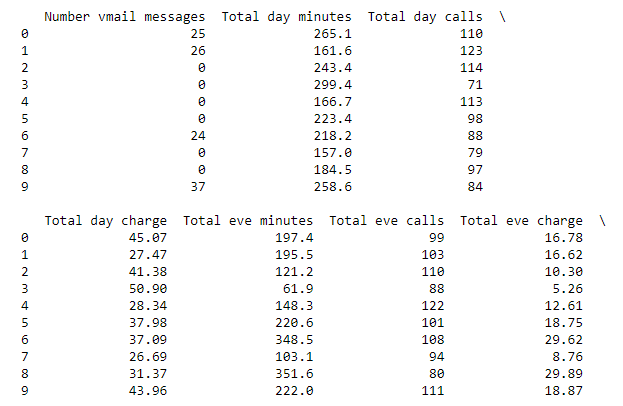
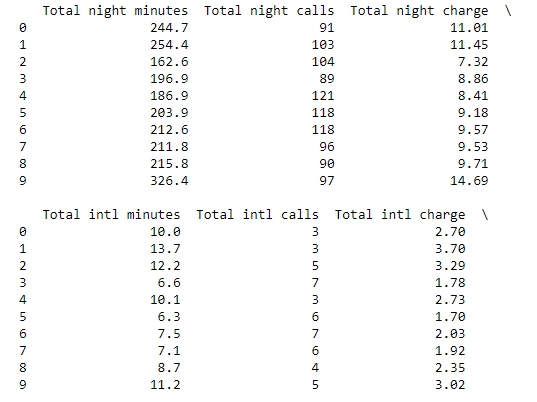
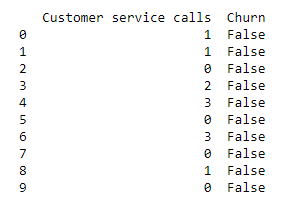


**Output:**



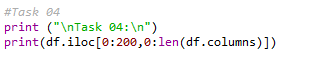
**Output at Jupyter**



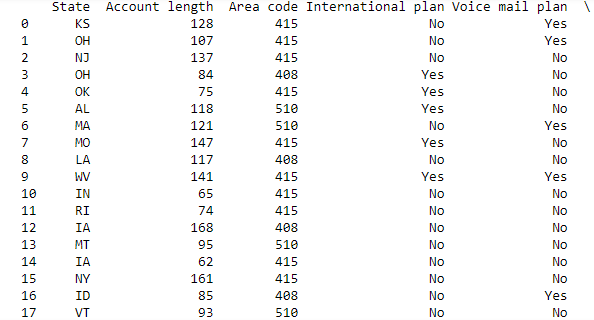
  

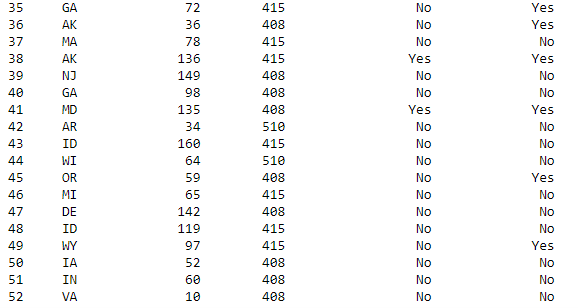
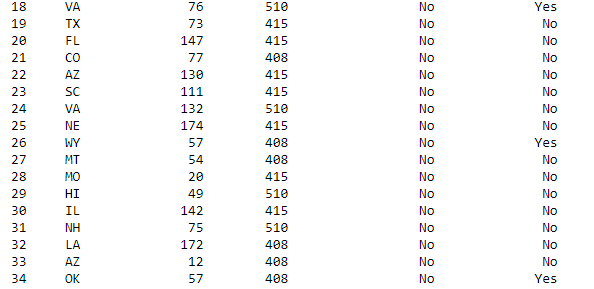
1. Write code to display whole dataframe with maximun 200 rows to be displayed in the output table.

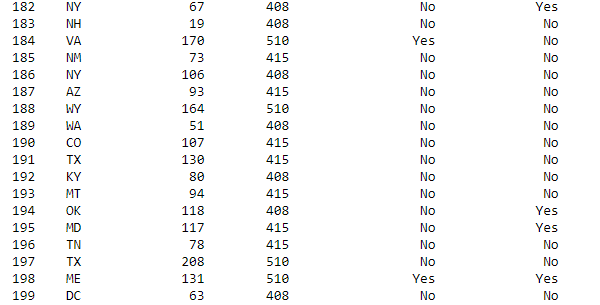
**Code:**



**Output:**





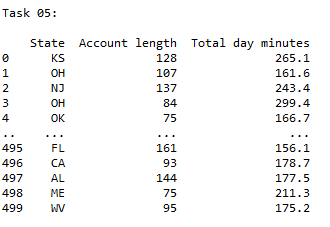


1. Write code to display state, account length, and today day minutes for first 500 clients.

**Code:**



**Output:**

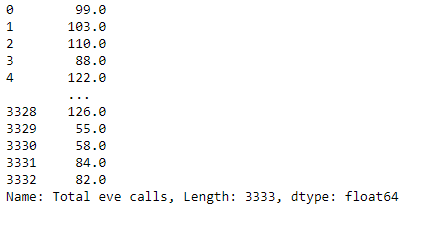


1. Write code to convert the data type of “Total eve calls” from int64 to **float64**.

**Code:**

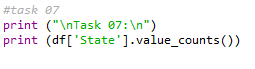


**Output:**

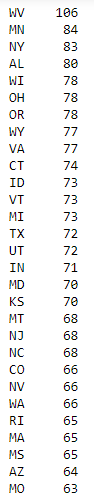


1. Write code to display value counts for state feature.

**Code:**



**Output:**

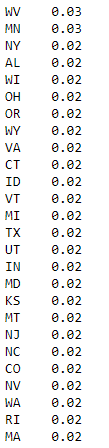


1. Write code to display fractional value counts for state feature.

**Code:**



**Output:**

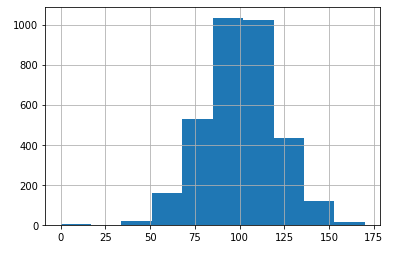


1. Write code to display a histogram for “Total eve calls”.

**Code:**



**Output:**

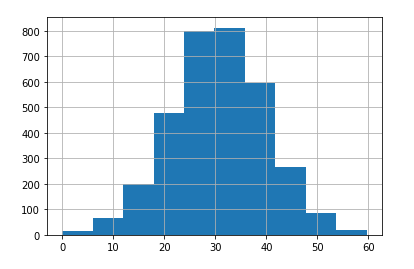


1. Write code to display a histogram for “Total day charge”.

**Code:**



**Output:**



**Submission Instructions:**

1. Solve your tasks in the provided notebook **“BI&BD\_SEN4210\_Lab4\_05\_04\_2020\_Studentid”** (Replace Studentid with your actual Student ID.
2. Submit your answers in this word file by renaming it in the format **“BI&BD\_SEN4210\_Lab{#}\_05\_04\_2020\_Student ID”** and uploading on **Moodle** in the appropriate submission link.
3. Along with this word file, submit your notebook containing the solution of lab tasks.

**Please conform to the naming convention of the file.**

**References:**

**[1]** Open Machine Learning Course. Web Access:<https://mlcourse.ai/>

**END OF LABORATORY**