Big Data Analytics (MPBA G517)

Lab Sheet - 5

Instructor: Dr Revendranath T
Teaching Assistants: Jainil Shah and Shashank Sharma

A.) Dataset titled 'filament.csv' is provided to you. Perform the following operations:

1. Create a Nested List of the Filament Data.
2. Create a Schema of the DataFrame.
3. Create a RDD of the Row Objects.
4. Create the DataFrame.
5. Print a Schema of the DataFrame.
6. Change the DataType of any given column of your choice.
7. Filter out the Data where BulbPower is 100 W.

B.) Using the same dataset, perform Exploratory Data Analysis on the DataFrame following the below mentioned steps:

1. Read Data from the CSV File and create a DataFrame.
2. Calculate Summary Statistics.
3. Count the frequency of distinct values in the FilamentType categorical column.
4. Count the frequency of distinct values in the BulbPower categorical column.
5. Counting the Frequency of Distinct Values in a Combination of FilamentType and BulbPower columns.

C.) Data file 'adultData.csv' is given. This is a simple CSV File with 15 columns. The following table describes all 15 columns.

Columns	**Description**
age	Age of person, continuous
workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt	Continuous
education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
Education-num	Continuous
Marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation	Tech-support, Craft-repair, Other-service, Sales, Execmanagerial, Prof-sepcialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship	Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex	Female, Male.
Capital-gain	Continuous
Capital-loss	Continuous
Hours-per week	Continuous
Native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, EI-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
Class (income)	>50K, <=50K

1. Create a DataFrame from the ‘adult.csv’ Datafile as per details given in the above table. Populate random values as per criteria given in each column.
2. Count the total number of records in the DataFrame.
3. Count the number of times that a salary is greater than $50,000 and the number of times it’s less than $50,000.
4. Perform summary statistics on the numeric columns age, capital-gain, capital-loss, and hours-per-week.
5. Find out the mean age of male and female workers from the data.
6. Find out whether a salary greater than $50,000 is more frequent for males or females.
7. Find the highest-paid job.