Logo

Description automatically generated

**MODULE 4**

**PRACTICE WITH APACHE SPARK**

Submitted By:

**HARSHIT GAUR**

MASTER OF PROFESSIONAL STUDIES IN ANALYTICS

ALY 6110 : DATA MANAGEMENT AND BIG DATA

OCTOBER 15, 2022

Submitted To: **PROF. DAYA RUDHRAMOORTHI**

# **Java and Apache Spark Installation**

1. Java was already installed in my system. I have updated it to the latest version available for MacOS.

Graphical user interface, text

Description automatically generated

1. After downloading the Apache Spark folder from its website, I have added the required paths for Apache Spark to work in the Bash-Profile of my system.

Graphical user interface, text, application, email

Description automatically generated

1. Updating the Bash-Profile of my system after making changes and running the ‘Spark Shell’ to check if changes have reflected or not.

Text

Description automatically generated

1. To verify if spark has installed properly, running some commands in Scala language.

Graphical user interface, text, application, email

Description automatically generated

1. For further verification with data, running some more commands to check it is working with data and its manipulation.

Text, email

Description automatically generated

1. Installing PySpark in the local machine to work in an editor with Spark and Python.

A screenshot of a computer

Description automatically generated

1. PySpark has been installed properly and successfully

A screenshot of a computer

Description automatically generated

1. PySpark library has also been properly imported into the IPYNB notebook after installation.

A screenshot of a computer

Description automatically generated

# **Data Set #1 (Census Data)**

## **Introduction**

We chose a dataset with 32560 rows and 14 different independent features. We would like to know if someone earns more than $50,000 per year. Because the data predicts just two outcomes (>50K or =50K), we know that this is a classification problem, and we will train classification models to predict the requisite outputs. The specifics of the features provided to us are listed below, and we will use them to train our categorization model.

* Age – The age of a person, which ranges from 17 to 90.
* Workclass - The sort of job to which a person is assigned.
* Fnlwgt — The weighted combination of features (an estimate of how many persons belong to this collection of features).
* Education — Highest level of education
* Education\_num — Number of years for which education was taken
* Marital Status — Represents the category assigned to a person based on his marriage status
* Occupation — Profession of a person
* Relationship — Relationship of the person in his family
* Race — Origin background of a person
* Sex — Gender of a person
* Native Country – The country of origin of a person.
* Income — The target variable that predicts whether or not an individual's income is greater than or less than $50,000.

**Question:**

**We want to know those people who make more than $50,000 each year.**

## **Analysis and Results**

1. Importing Spark Libraries, Visualization Libraries, Pandas-NumPy (Data Manipulation) Libraries.

Graphical user interface, text, application

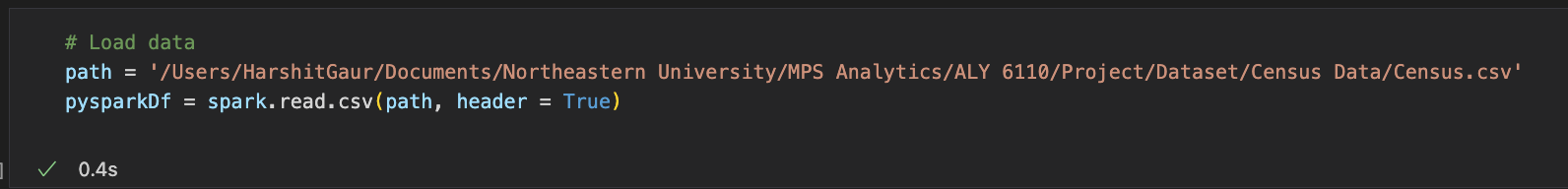
Description automatically generated

1. Starting “Apache Spark” connection with Python.

Graphical user interface, text, website

Description automatically generated

1. Loading the Dataset #1 (Census Data) using PySpark.



1. Displaying and checking the Schema of the data.

Text

Description automatically generated

1. To check the first few records of the Census data.

A screenshot of a computer screen

Description automatically generated with medium confidence

1. The total count of records in the Census dataset is 32,561.

Graphical user interface, text, application

Description automatically generated

1. To get the statistical properties of some of the numeric variables (features) of the dataset.

Graphical user interface, text

Description automatically generated with medium confidence

1. Next step in the analysis part is to check whether the dataset has missing, NA, null values.
   1. To check for missing values.

A screenshot of a computer

Description automatically generated with medium confidence

|  |  |
| --- | --- |
| b. To check for NULL values.  Text  Description automatically generated | c. To check for values with ?  Text  Description automatically generated |

1. Checking the correlation between numeric variables of the dataset, a correlation matrix was generated.

Text

Description automatically generated

1. In order to display the correlation in a chart, a Heat Map was also created.

Table

Description automatically generated

*We can check that the variables are not significantly correlated with each other in this dataset.*

1. A Histogram was plotted to understand the distribution of ‘Age’ attribute in Census.

Text

Description automatically generated

Chart, histogram

Description automatically generated

*The percentage of people in the data set is very high in the age bracket of 22 – 42.*

1. To address the understanding of ‘Education Type’ in the data set, a bar plot was generated.

Text

Description automatically generated

Chart, histogram

Description automatically generated

## **Insights**

* The census data contains a very high percentage of people who have education in 3 categories *(****HS-Grad, Bachelors, Some-College****).*
* To find out those people who are and who will be earning more than $50,000 per year will depend heavily on some of the factors including but not limited to:
  + Education
  + Occupation
  + Age

# **Data Set #2 (Car Dekho Sales)**

## **Introduction**

CarDekho.com is India's leading automobile search engine, enabling customers in selecting the ideal vehicle for them. Its website and app offer a wealth of automotive content, including expert reviews, full specifications and prices, comparisons, and videos and images of all automobile brands and models available in India.

We have sales figures for every car sold between 1983 and 2020. We will analyze the acquired data to assist them in growing their business, gaining and retaining consumers, and outperforming their competitors.

The data set contains 8128 data points and 13 features pertaining to:

* Car specifications include the vehicle's name, transmission, fuel type, seating capacity, and year of manufacturing.
* Engine details include mileage, engine type, torque, and maximum power in BHP.
* The selling price and the amount of kilometers driven by the car are among the details of the transaction.

**Question:**

**Identify the right prospects at the right time.**

**Boost client loyalty**

## **Analysis and Results**

1. Loading the Dataset #1 (CarDekho Sales Data) using PySpark.

**Graphical user interface, text

Description automatically generated**

1. Displaying and checking the Schema of the data.

Text

Description automatically generated

1. To check the first few records of the Census data.

Graphical user interface, text

Description automatically generated

1. The total count of records in the Census dataset is 8,128.

Graphical user interface, text, application

Description automatically generated

1. Finding out the Statistical Properties of some of the numeric variables of the data set.

Text

Description automatically generated with low confidence

1. To check for missing values.

Text

Description automatically generated

1. To check for NULL values.

Text

Description automatically generated

1. Checking the correlation between numeric variables of the dataset, a correlation matrix was generated.

Text

Description automatically generated

1. In order to display the correlation in a chart, a Heat Map was also created.

Graphical user interface, website

Description automatically generated with medium confidence

*We can check that the some variables are significantly correlated with each other in this dataset.*

* *Year is around 40% correlated with Selling Price and Km-Driven.*

1. A Histogram was plotted to understand the distribution of ‘Seats’ attribute in the CarDekho Sales data.

Text

Description automatically generated

Chart, histogram

Description automatically generated

## **Insights**

* A very high percentage of cars were sold during the years 2012-2017.
* This might be due to influx of modern transmission techniques (mainly, Automatic Transmission) in this industry.

# **Data Set #3 for JOINS (Countries-Continent)**

## **Introduction**

This data set contains the linking of countries with their respective continents in order to aid us joining our first data set of ***Census Data*** with this ***Countries-Continent Data*** to retrieve information regarding the Continent to which each data point (citizen) belong to.

## **Analysis and Results**

1. Loading the Dataset #3 (Countries-Continent Data) using PySpark.

Graphical user interface, text, website

Description automatically generated

1. Displaying the records of the data set.

Graphical user interface, text

Description automatically generated

1. Performing INNER JOIN to join 2 data sets and retrieve information.

Graphical user interface, text, application

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

# **References**

*UCI Machine Learning Repository: Census Income Data Set*. (n.d.). Retrieved October 16, 2022, from [*https://archive.ics.uci.edu/ml/datasets/census+income*](https://archive.ics.uci.edu/ml/datasets/census+income)

*Vehicle dataset. (2020, October 24). Kaggle. <https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho>*

*GitHub - dbouquin/IS\_608: Coursework for 608 Knowledge and Visual Analytics, MS in Data Analytics at CUNY SPS*. (n.d.). GitHub. Retrieved October 16, 2022, from https://github.com/dbouquin/IS\_608/