Assignment Content

Top of Form

**BDAT 1009 – PySpark**

**Lab 3: PySpark**

|  |  |
| --- | --- |
| **Student Name** | **Student Number** |
| Namppd-e-Sehar | 200474523 |

Colab: <https://colab.research.google.com/>

Automobile Data: <https://www.kaggle.com/datasets/toramky/automobile-dataset>

Tutorial: https://medium.com/linkit-intecs/pyspark-with-google-colab-d964fd693ca7

Follow the above tutorial and create your own Colab Notebook (Need free Google account).

Download and load the automobile data to your Google Drive.

Import the data into a PySpark Dataframe. Perform the following tasks. Provide brief desc on what you did for each step, your code and screenshot.

A screenshot of a computer

Description automatically generated

1. Load Data Automobile data into a PySpark DF. After running ‘head’ command.

2) Run PySpark command to get Count. What is the count?

3) Run PySpark command to get the Schema.

A screenshot of a computer

Description automatically generated

1. Run PySpark command Remove Nulls.

A screenshot of a computer

Description automatically generated

1. Filtering: Let’s get the cars with highway mpg more than 23.

A screenshot of a computer

Description automatically generated

1. Multiple Conditions: Let’s get cars with horsepower > 80 and curb-weight more than 2000.

A screenshot of a computer

Description automatically generated

1. Select only Audi make cars.

A screenshot of a computer

Description automatically generated

1. Get count of cars grouping by Make and Drive-Wheels.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Provide both Notebook file and this document with Your code, screenshot and brief description on what you did for each task above.

Bottom of Form