DSCI 5350 Spark Activity-4 Satyasriharsha Buddha

Spark ML pipeline-Activity 4

Import the File

A screenshot of a computer

Description automatically generated

Path - dbfs:/FileStore/shared\_uploads/sriharshabuddha@my.unt.edu/Diamonds-1.csv

1)Load the dataset as a data frame

A screenshot of a computer

Description automatically generated with medium confidence

2) Print the schema of the data set.

A picture containing text, screenshot

Description automatically generated

3)Count the number of records and display first 10 records.

A screenshot of a computer

Description automatically generated with medium confidence

4) By default, the data type of price (Dependent variable) is taken as integer. We have to change that to double.

A screenshot of a computer

Description automatically generated with medium confidence

5)Display the top 5 records and verify the price column

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

Categorical feature- cut, color, clarity

Continuous feature-depth, x, y, z

6)Show distinct values in categorical features

A screenshot of a computer program

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

7) Display the distinct values in cut. Click drop down in chart options and select bar chart

Select plot options and do the following settings

Value=<id> ; Series groupings=’cut’;Aggregation=COUNT

A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated with low confidence

8) Visualize the continuous features

Depth-

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

Column X

A screenshot of a chat

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

Column Y

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated with low confidence

Column Z

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Column Carat

A picture containing screenshot, text

Description automatically generated

A screenshot of a graph

Description automatically generated with medium confidence

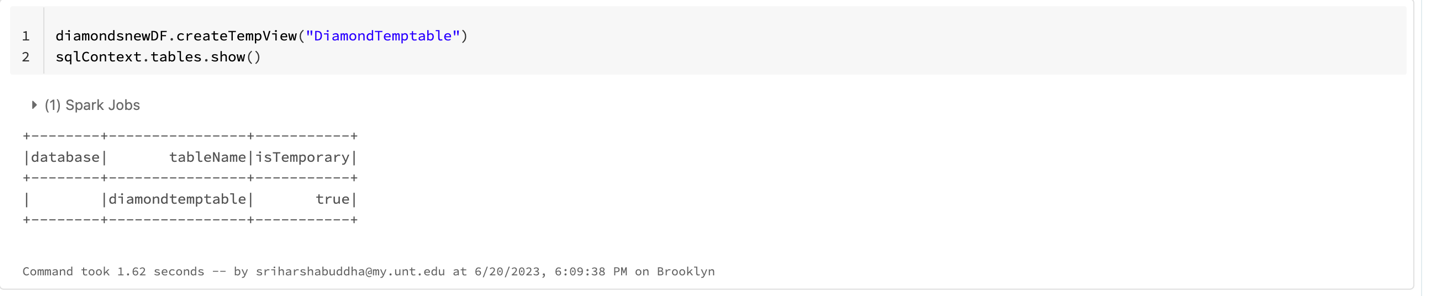
Most of the diamonds are small . The above visualization shows a skewed distribution.

9) Using interactive SQL to explore the data

A screenshot of a computer

Description automatically generated with medium confidence

10) Create a temporary view



11) Write sql query to fetch data from the view

A screenshot of a computer

Description automatically generated with medium confidence

10) Select all the values from the temp table where clarity starts with ‘I’

A screenshot of a computer

Description automatically generated with medium confidence

11) Execute Aggregate function to get the average price of the diamonds

A screenshot of a chat

Description automatically generated with medium confidence

12)Display carat, clarity, price of top 5 records sorted by carat field

A screenshot of a computer

Description automatically generated with low confidence