PySpark Assignment

1. Using the PySpark, answer the following questions.

Copy your code into a Word document. Submit your document to Moodle.

I. Load the loan_200k.csv into a Spark Dataframe

II. Display/show the first 10 rows

III. How many records are in the Dataframe?

IV. Display descriptive stats using the describe function

V. We are feeling generous. Create a new column called ‘new_amnt’ that has *floor* of the ‘payment’. Floor is already a function in pyspark.sql.functions but you need to import it in order to use it.

VI. Load the DF that you just created into a Spark temporary view (so that we can query the table using SparkSQL)

VII. Use SparkSQL to query the temporary view that you just loaded in the previous step: for each of the purposes of the loans, show/display what is the average income? – Order your results by descending average income.

VIII. Use SparkSQL to query the temporary view that you loaded in the step VI: show/display how many records (observations) defaulted, and how many did NOT defaulted?

IX. Use SparkSQL to query the temporary view that you loaded in the step VI: for each of the following states: NC, CA, MA, TX, FL, show/display what is average payment?