**Instructions:**

Evaluate the homework against the outlined criteria in the below rubric, assigning a rating to each criterion. Add points earned across all criteria and convert the total points to a letter grade, assigning a “+” or “-” letter grade designation at your discretion.

| | | | | | |
|---|---|---|---|---|---|
| A (+/-) | 70+ | C (+/-) | 30-49 | F (+/-) | <10 |
| B (+/-) | 50-69 | D (+/-) | 10-29 | | |

**Notes:**

The deployed assignment utilizes **AWS** and the **PySpark** library to complete 1 of 2 levels of challenges. The source code should also be deployed to **Github** or **Gitlab**.

**Rubric for Big Data - Level 1:**

| | **Mastery<br>20 points** | **Approaching Mastery<br>15 points** | **Progressing<br>10 points** | **Emerging<br>5-0 points** | **Incomplete** |
|---|---|---|---|---|---|
| **Extract** | In both notebooks the student did all of the following:<br><br>✓ Connects to and loads in datasets from AWS to dataframes using pyspark<br>✓ Correctly handles the header and has column names as the first row<br>✓ Ensures the data is retrieved by outputting the head of the dataframe<br>✓ Discovers the size of the dataframe by outputting the number of rows in it | In both notebooks the student did 3 of the following:<br><br>✓ Connects to and loads in datasets from AWS to dataframes using pyspark<br>✓ Correctly handles the header and has column names as the first row<br>✓ Ensures the data is retrieved by outputting the head of the dataframe<br>✓ Discovers the size of the dataframe by outputting the number of rows in it | In both notebooks the student did 2l of the following:<br><br>✓ Connects to and loads in datasets from AWS to dataframes using pyspark<br>✓ Correctly handles the header and has column names as the first row<br>✓ Ensures the data is retrieved by outputting the head of the dataframe<br>✓ Discovers the size of the dataframe by outputting the number of rows in it | In both notebooks the student did 0-1 of the following:<br><br>✓ Connects to and loads in datasets from AWS to dataframes using pyspark<br>✓ Correctly handles the header and has column names as the first row<br>✓ Ensures the data is retrieved by outputting the head of the dataframe<br>✓ Discovers the size of the dataframe by outputting the number of rows in it | No submission was received<br><br>-OR-<br><br>Submission was empty or blank<br><br>-OR-<br><br>Submission contains evidence of academic dishonesty |
| **Transform<br>&<br>Load** | Student does all of the following with the dataframes:<br><br>**Transform**<br>✓ Removed duplicate rows<br>✓ Kept and renamed only | Student does 3 of the following with the dataframes:<br><br>**Transform**<br>✓ Removed duplicate rows<br>✓ Kept and renamed only | Student does 2 of the following with the dataframes:<br><br>**Transform**<br>✓ Removed duplicate rows<br>✓ Kept and renamed only | Student does 0-1 of the following with the dataframes:<br><br>**Transform**<br>✓ Removed duplicate rows<br>✓ Kept and renamed only | |

| | | | | | |
|---|---|---|---|---|---|
| | necessary columns to match the current database table schema<br>✓ Matched dataframe column types with the database column types<br><br>**Load**<br>✓ Successfully pushed dataframes to AWS | necessary columns to match the current database table schema<br>✓ Matched dataframe column types with the database column types<br><br>**Load**<br>✓ Successfully pushed dataframes to AWS | necessary columns to match the current database table schema<br>✓ Matched dataframe column types with the database column types<br><br>**Load**<br>✓ Successfully pushed dataframes to AWS | necessary columns to match the current database table schema<br>✓ Matched dataframe column types with the database column types<br><br>**Load**<br>✓ Successfully pushed dataframes to AWS | |

**Rubric for Big Data - Level 2:**

| | **Mastery**<br>**20 points** | **Approaching Mastery**<br>**15 points** | **Progressing**<br>**10 points** | **Emerging**<br>**5-0 points** | **Incomplete** |
|---|---|---|---|---|---|
| **Extract & Transform** | The student did all of the following to extract and clean the data:<br><br>✓ Connected to and loaded in datasets from AWS to dataframes using Spark<br>✓ Removed any unnecessary columns<br>✓ Dropped rows with null values<br>✓ Dropped duplicated rows | The student did 3 of the following to extract and clean the data:<br><br>✓ Connected to and loaded in datasets from AWS to dataframes using Spark<br>✓ Removed any unnecessary columns<br>✓ Dropped rows with null values<br>✓ Dropped duplicated rows | The student did 2 of the following to extract and clean the data:<br><br>✓ Connected to and loaded in datasets from AWS to dataframes using Spark<br>✓ Removed any unnecessary columns<br>✓ Dropped rows with null values<br>✓ Dropped duplicated rows | The student did 0-1 of the following to extract and clean the data:<br><br>✓ Connected to and loaded in datasets from AWS to dataframes using Spark<br>✓ Removed any unnecessary columns<br>✓ Dropped rows with null values<br>✓ Dropped duplicated rows | No submission was received<br><br>-OR-<br><br>Submission was empty or blank<br><br>-OR-<br><br>Submission contains evidence of academic dishonesty |
| **Analysis** | Student does all of the following to analyze if “vine” reviews are trustworthy:<br><br>✓ Splits the reviews between vine (paid) and non-vine (unpaid)<br>✓ Compares metrics between vine and non-vine reviews such as, but not limited to:<br>`Number of reviews`<br>`Number of 5-star reviews`<br>`Average Rating`<br>`Number of helpful votes`<br><br>✓ Comes up with a conclusion on the trustworthiness of vine reviews with data to back up their claim | Student does all of the following to analyze if “vine” reviews are trustworthy:<br><br>✓ Splits the reviews between vine (paid) and non-vine (unpaid)<br>✓ Compares metrics between vine and non-vine reviews such as, but not limited to:<br>`Number of reviews`<br>`Number of 5-star reviews`<br>`Average Rating`<br>`Number of helpful votes`<br><br>✓ Comes up with an **invalid** conclusion on the trustworthiness of vine reviews or **does not provide** data to back up their claim | Student does 2 of the following to analyze if “vine” reviews are trustworthy:<br><br>✓ Splits the reviews between vine (paid) and non-vine (unpaid)<br>✓ Compares metrics between vine and non-vine reviews such as, but not limited to:<br>`Number of reviews`<br>`Number of 5-star reviews`<br>`Average Rating`<br>`Number of helpful votes`<br><br>✓ Comes up with a conclusion on the trustworthiness of vine reviews with data to back up their claim | Student does 0-1 of the following to analyze if “vine” reviews are trustworthy:<br><br>✓ Splits the reviews between vine (paid) and non-vine (unpaid)<br>✓ Compares metrics between vine and non-vine reviews such as, but not limited to:<br>`Number of reviews`<br>`Number of 5-star reviews`<br>`Average Rating`<br>`Number of helpful votes`<br><br>✓ Comes up with a conclusion on the trustworthiness of vine reviews with data to back up their claim | |